METODI STATISTICI E COMPUTAZIONALI

Stefania Spagnolo

Dipartimento di Matematica e Fisica, Univ. del Salento

ANALISI DATI

Lezione 5

ANALISI DATI

Tipologie di Osservabili

Quantitative (hard)

- Si può associare un valore numerico ed unità di misura
- Es.: La temperatura dell'acqua in K. Le osservabili quantitative sono la base della scienza galileiana e delle cosiddette "hard sciences": le discipline basate su osservabili rigorose connesse tra loro da modelli matematici.

Qualitative (soft)

- Etichette, descrittori, categorie
- Generalmente sono espresse verbalmente ("Il corpo è sferico, il soggetto ha un carattere irritabile").
- Osservabili difficilmente standardizzabili e/o riproducibili (es. analisi sensoriale)
- Fuzzy logics: tentativo di rendere quantitative delle osservabili espresse verbalmente

ANALISI DATI

Tipologie di Osservabili

Osservabili **discrete**

Sono associati a processi che possono assumere solo un insieme limitato di valori predefiniti (es. facce di un dado, stato di spin di una particella)

Osservabili **continue**

Sono associati a processi che possono dare origine a un insieme continuo di valori possibili. Tali valori possono essere limitati o illimitati

(es. temperatura di un liquido, velocità di una particella)

I limiti strumentali possono dar luogo a discretizzazioni

Esempio conversione Analogico-Digitale

ANALISI DATI

In termini statistici

- Un ***insieme S contenente tutti i possibili risultati di un esperimento è detto spazio campione*** e ciascun risultato è detto elemento o punto di S. ***Un evento (A) in S è costituito da 1 o più elementi di S***. L'insieme di tutti gli eventi in S è anche detto campo di Borel
- Uno spazio campione può essere finito o infinito, continuo o discreto.
  - es.: lancio di un dado => spazio discreto finito
  - serie storica a lotto => spazio discreto infinito
  - misura di una lunghezza => spazio continuo
- La ***variabile aleatoria mi permette di associare ad un evento dello spazio di Borel un valore numerico.***
  - Esempi: La temperatura dell'acqua è 400 K; Indico con 1 i corpi sferici.

$$X : \mathrm{B} \rightarrow \Re$$

- Variabili ***aleatorie discrete e continue finite o infinite***
-

$$X(w_i) = x_i \qquad w_i \in \mathrm{B}, x_i \in \Re$$

In termini statistici

- Per indicare il valore assunto dalla variabile aleatoria che identifica il risultato numerico di un esperimento si usa il termine ***dato*** o ***misura***.

- Per misure sperimentali occorre ricordare che:
  - Esiste sempre una ***calibrazione*** che lega il dato al valore dell'osservabile. Lo strumento, qualunque esso sia, è calibrato per effettuare la misura.
  - La tecnica di misura usata può non essere selettiva. Una misura si dice ***selettiva*** se esiste una ***corrispondenza biunivoca tra ingresso dello strumento e risultato della misura***.
  - ***L'osservabile può avere una sua legge e/o una sua distribuzione di probabilità.*** In questo caso è possibile solo l'estrazione di un campione e la valutazione delle proprietà della distribuzione di probabilità che è responsabile del processo.
  - ***Il dato è soggetto a processi stocastici*** legati alla tecnica utilizzata nell'esperimento, al tipo di strumenti usati e all'operato degli sperimentatori. La misura è caratterizzata da un valore e da un errore.

Calibrazione

- Quasi sempre la singola misura è il risultato di un'azione che richiede l'utilizzo di uno o più strumenti di misura.
- Gli strumenti di misura danno una risposta che è legata all'osservabile che si vuole valutare attraverso una calibrazione

Es.: Misura di una temperatura con un termometro ad alcool.
- Osservabile (T): temperatura
- Risposta strumento (L): dilatazione lineare su scala graduata

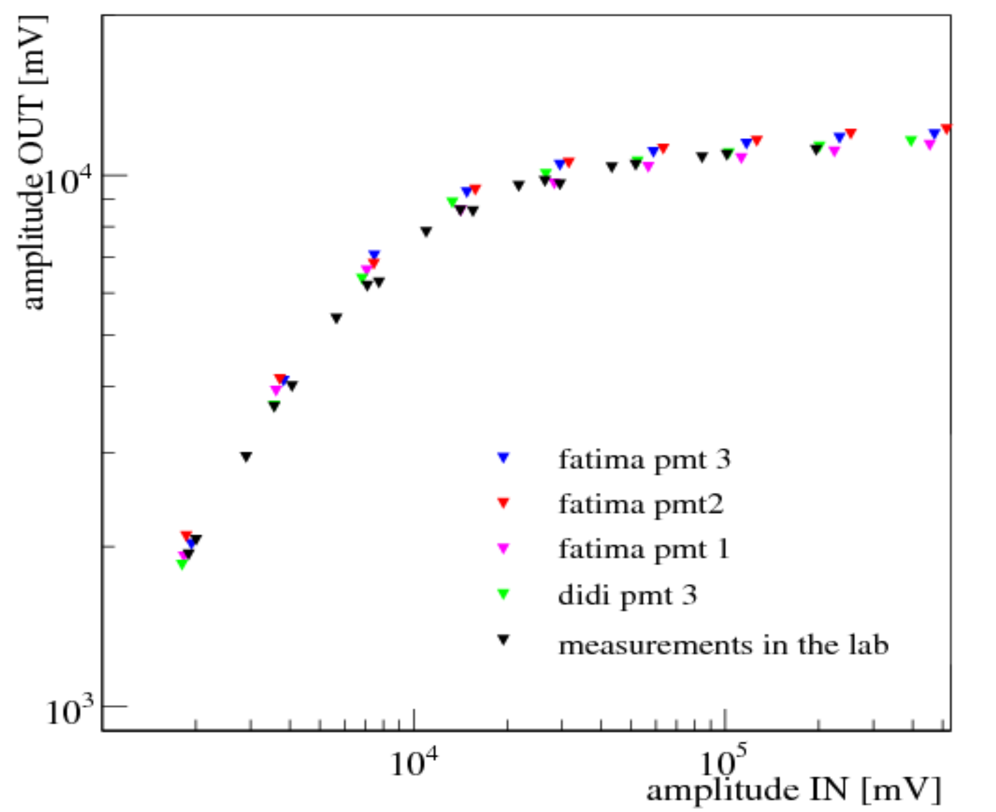
$$L = k \cdot T$$

k costante di calibrazione

Questo è un esempio di variabile aleatoria **selettiva** associata ad un osservabile. La calibrazione è **lineare**.

- Non sempre, però, si è in queste condizioni. Si possono avere funzioni di calibrazione non lineari e/o variabili aleatorie non selettive.

ANALISI DATI

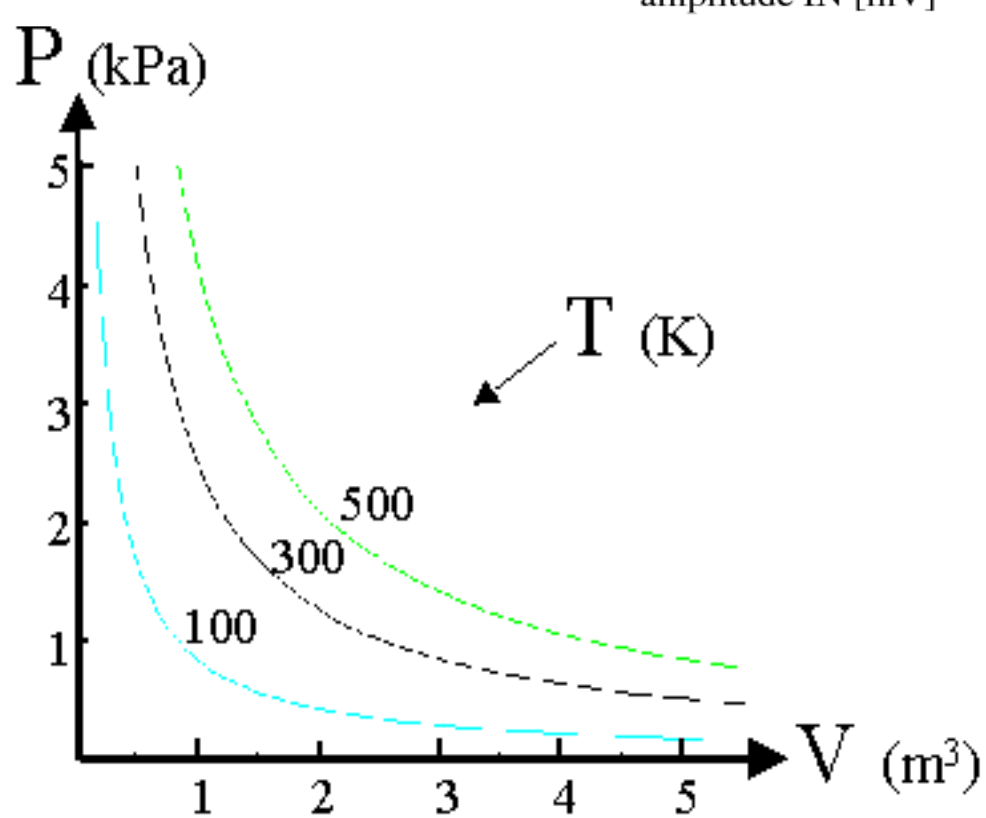


Misura selettiva e non selettiva

Curva di risposta di un fotomoltiplicatore.
Misura ***selettiva:***

- ad ogni tensione di ingresso corrisponde una e una sola ampiezza del segnale in tensione di uscita.
- sono chiaramente distinguibili una regione di linearità (sino a circa 4-5 V) e una regione di non linearità.

La funzione di ***calibrazione è non lineare***.
L'errore associato alla misura non è costante.



Equazione dei gas perfetti.
Per determinare lo stato del gas occorre misurare due grandezze. La sola misura della pressione non mi permette di determinare il volume. P **non è una misura selettiva** di V.

$$P = \frac{nRT}{V}$$

Errore di Misura

- **Il risultato di una misura è una variabile aleatoria** in quanto al **processo di misura** sono associate fluttuazioni dovuta a varie cause che rendono la misura non esattamente riproducibile. La misura è riproducibile sono in termini statistici.

$$Y = k \cdot X + \varepsilon$$

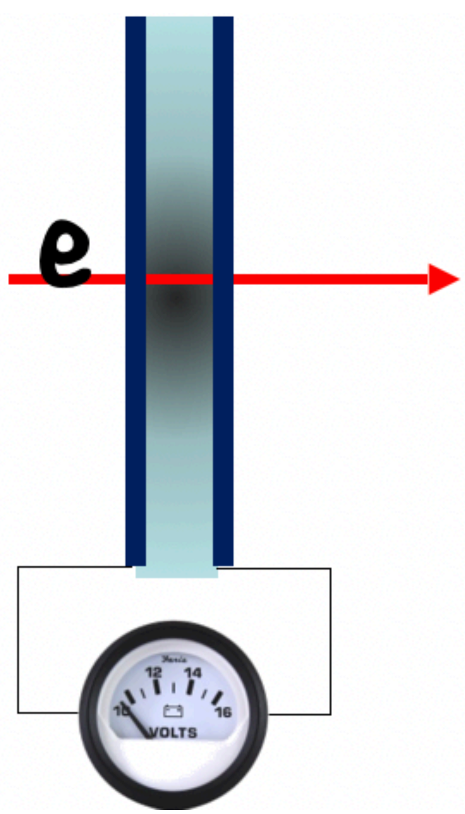
- Y risposta dello strumento, X osservabile da misurare, ε errore della misura. Tipicamente si assume che gli errori statistici fluttuino in maniera gaussiana

Teorema del Limite Centrale

"se si ha una somma di variabili aleatorie indipendenti e identicamente distribuite con media μ e varianza σ^2, allora indipendentemente dalla forma distributiva di partenza, al tendere della dimensione campionaria a infinito la somma tende a distribuirsi come una variabile casuale normale"

Processi fisici stocastici

- L'***osservabile*** che si vuole valutare può essere a sua volta frutto di un processo stocastico con una sua distribuzione di probabilità.
- Esempio
  - una particella carica che attraversa un sottile strato di materia rilascia nel mezzo una frazione della sua energia. Il processo alla base di questo fenomeno è un processo stocastico. ***La distribuzione di probabilità*** dell'energia ceduta al mezzo è una distribuzione di ***Landau***.
  - Supponiamo di voler misurare questa perdita attraverso un rivelatore che mi produce una d.d.p. direttamente proporzionale alla quantità di energia rilasciata.



$$\Delta V = k \cdot \Delta E + \varepsilon$$

ΔE Segue una distribuzione di Landau

ε Segue una distribuzione di Gauss

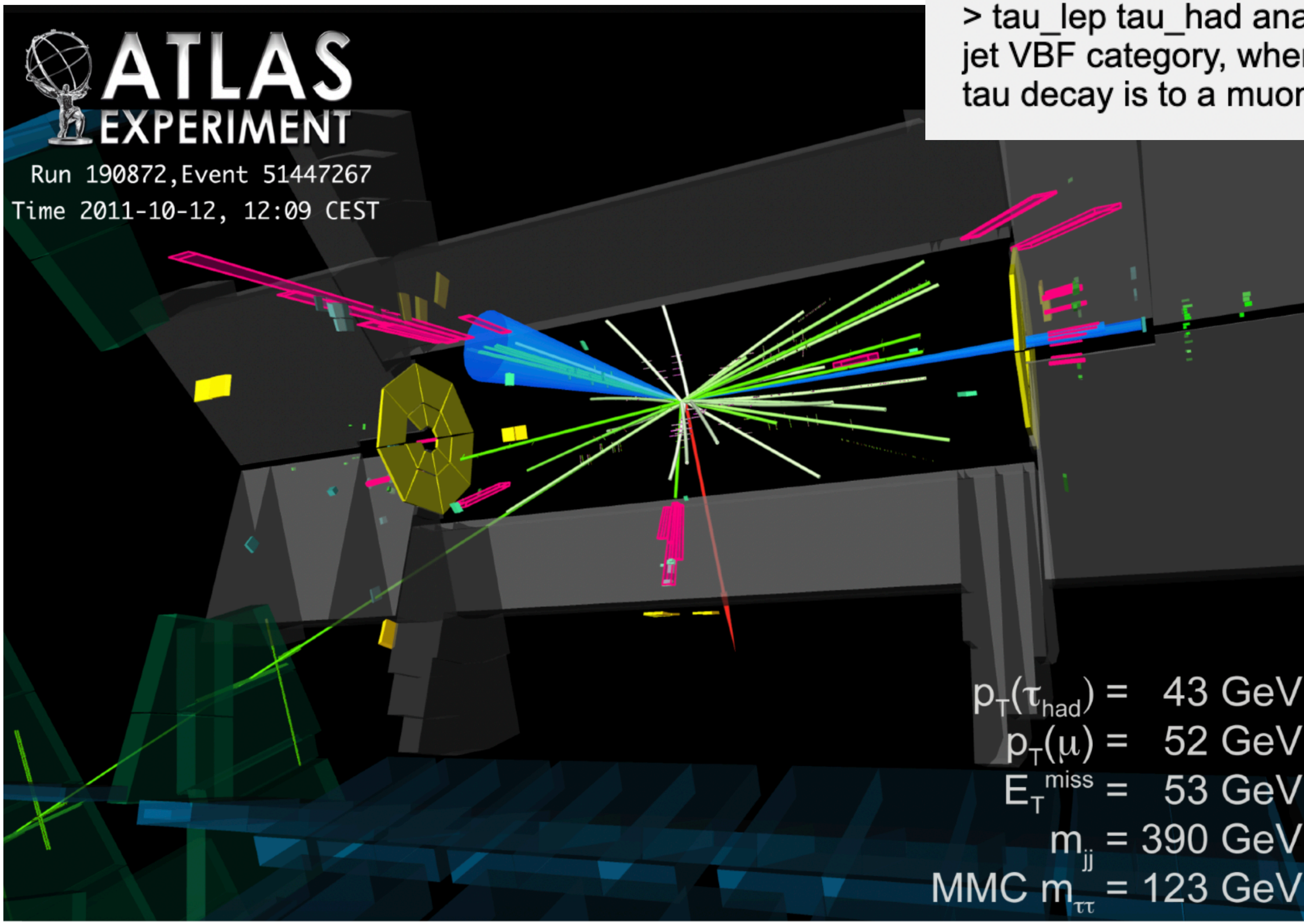
ANALISI DATI

- Molti procedimenti sperimentali tendono a produrre ***dati univariati*** in cui cioè il dato sperimentale dipende da una sola grandezza
  - Misura di una singola grandezza incognita
- Anche in questo caso "semplice" la determinazione del risultato richiede una serie di accortezze:
  - Attenzione alle possibili interferenze con la misura
  - Mantenere costanti le condizioni in cui la misura è eseguita
  - Preparare accuratamente l'esperimento per isolare la grandezza da misurare

  - Nonostante ciò, molti fenomeni sono intrinsecamente complessi e modellabili solo considerando una molteplicità di indicatori

- Alcuni apparati sperimentali producono una grande quantità di dati evento

ANALISI DATI

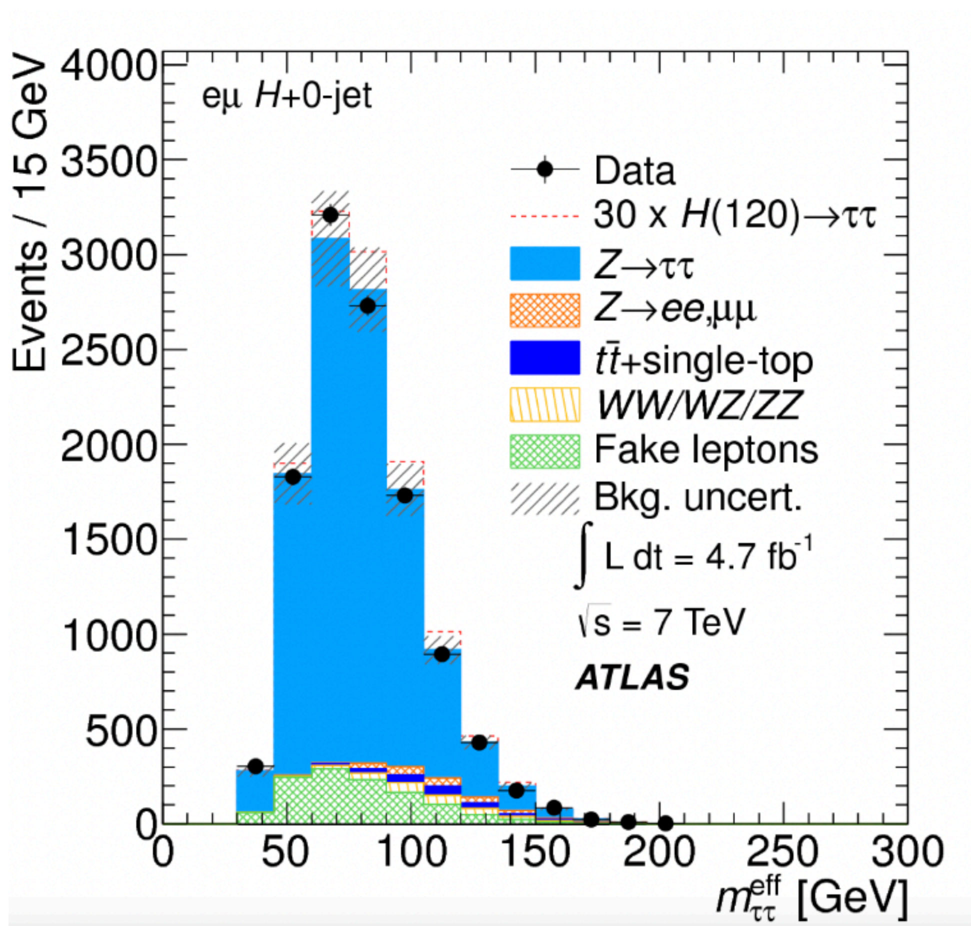


Display of an event selected by the H -> tau_lep tau_had analysis in the H+2-jet VBF category, where the leptonic tau decay is to a muon. (*figaux_17*)

ANALISI DATI

- Col termine ***analisi multivariata*** si indica quell'insieme di metodi statistici usati per analizzare simultaneamente piu' caratteri.

- In tutte le analisi statistiche multivariate il materiale grezzo e' costituito da **un certo numero di *eventi* (osservazioni) che si vogliono studiare simultaneamente**.



-

ANALISI DATI

- Col termine ***analisi multivariata*** si indica quell'insieme di metodi statistici usati per analizzare simultaneamente piu' caratteri.
- In tutte le analisi statistiche multivariate il materiale grezzo e' costituito da un certo numero di ***eventi*** (osservazioni) che si vogliono studiare simultaneamente.
- In altri casi si cercano correlazioni relazioni funzionali tra gruppi di variabili (***regressione***)
- In altri casi ancora si può essere interessati a ridurre le dimensioni della variabile multipla considerata (identificazione delle ***componenti principali***).
- In alcuni casi l'obiettivo dell'analisi e' semplicemente quello di classificare gli eventi sulla base di tutte le variabili considerate (***classificazione***).
- L'analisi multivariata è comunemente usata in fisica e inconsciamente ne avete già fatto uso. La regressione è un esempio di analisi multivariata.

ANALISI DATI

- Le precedenti osservazioni hanno evidenziato il legame che esiste tra il concetto di ***misura*** e la ***statistica***.
- L'aleatorietà del processo di misura dovuta alla presenza di un errore di misura rende indispensabile lo studio della statistica.

- Nel corso delle prossime lezioni affronteremo rapidamente nell'ordine:
  - **Il concetto di probabilità nelle sue varie formulazioni**
  - **Le variabili aleatorie** come grandezze in grado di quantificare un evento casuale
  - **La teoria dei campioni** e il suo legame con il processo si misura.

- Durante tutto il corso faremo spesso ricorso a simulazione di processi mediante tecniche Monte Carlo per mettere in pratica o "sperimentare" quanto visto in teoria.

- Affronteremo, infine, tecniche di analisi dati più specifiche e metodi numerici utili per risolvere problemi fisici.

Richiami di Calcolo Combinatorio e Probabilità

CALCOLO COMBINATORIO E PROBABILITÀ

- ***Un insieme S contenente tutti i possibili risultati di un esperimento*** è detto ***spazio campione*** e ciascun risultato è detto elemento o punto di S.
- Un ***evento (A) in S*** è costituito da 1 o più elementi di S. ***L'insieme di tutti gli eventi in S è anche detto campo di Borel***

- Uno spazio campione può essere finito o infinito, continuo o discreto
- Esempi:
  - lancio di un dado == spazio discreto finito
  - serie storica a lotto == spazio discreto infinito
  - misura di una lunghezza == spazio continuo
- Esiste un evidente parallelismo tra l'**algebra degli eventi** e quella degli ***insiemi***

Unione	$A \cup B$	evento A oppure B
Intersezione	$A \cap B$	sia l'evento A che quello B
Complementare	$\overline{A}$	l'evento non A
Differenza	$A - B$	l'evento A ma non B

CALCOLO COMBINATORIO E PROBABILITÀ

- Sommario

Proprietà delle operazioni insiemistiche:

$A \cup B = B \cup A \qquad A \cap B = B \cap A$ Proprietà commutativa

$A \cup (B \cup C) = (A \cup B) \cup C \qquad A \cap (B \cap C) = (A \cap B) \cap C$ associativa

$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$

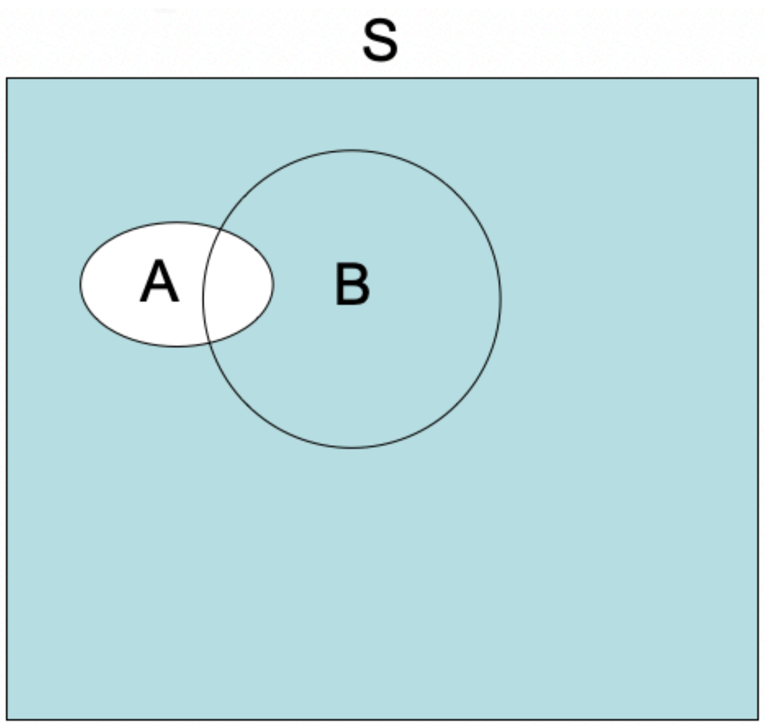
$A \cap (B \cup C) = (A \cup B) \cup (A \cap C)$

Proprietà distributiva

$\overline{A} \cup \overline{B} = \overline{(A \cap B)}$ Legge di De Morgan

$\overline{A} \cap \overline{B} = \overline{(A \cup B)}$ Legge di De Morgan

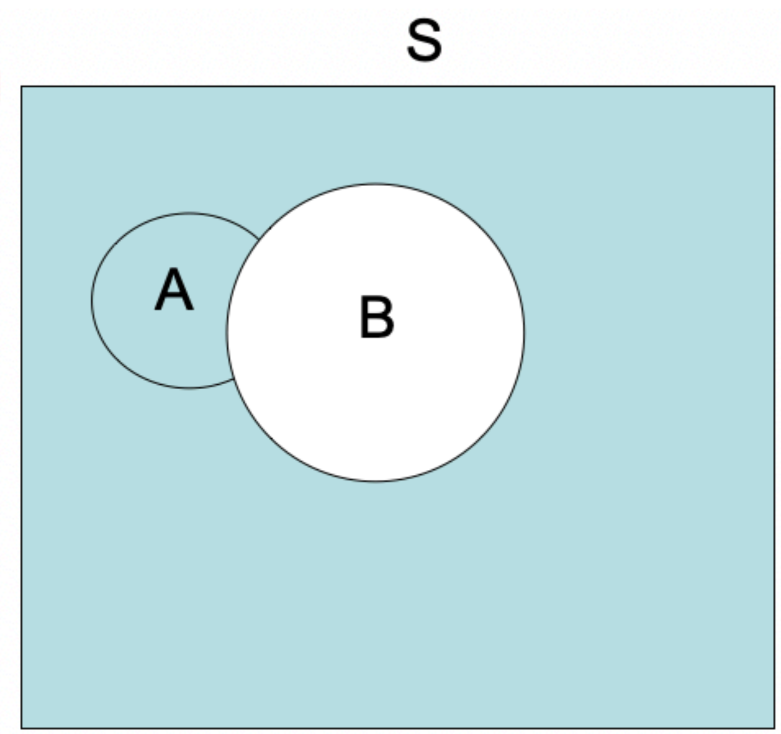
CALCOLO COMBINATORIO E PROBABILITÀ



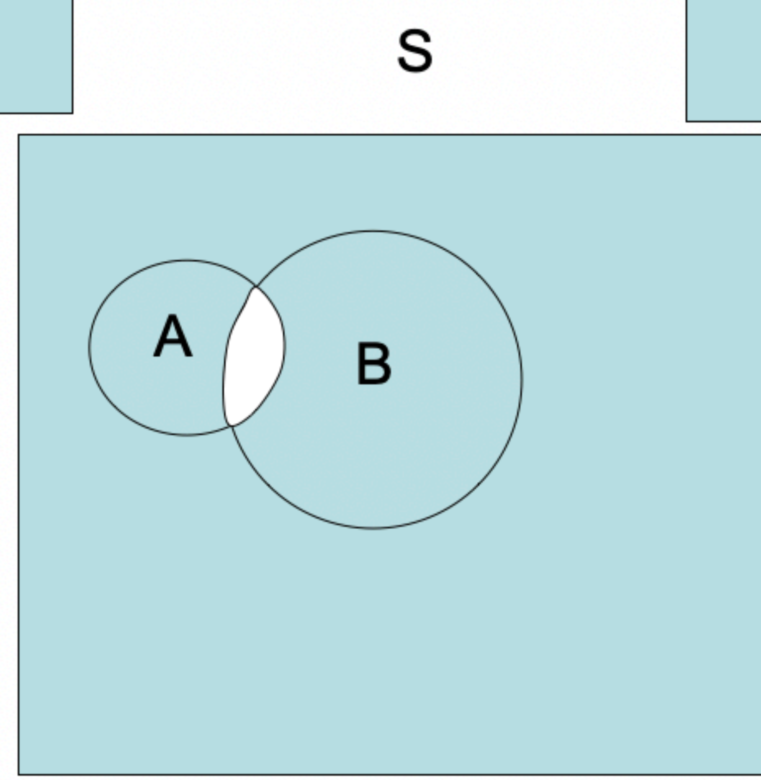
Area azzurra $\overline{A}$

Legge di De Morgan

$$\overline{A \cap B}$$



Area azzurra $\overline{B}$



Area azzurra $\overline{A} \cup \overline{B}$

$$\overline{A} \cup \overline{B} = \overline{(A \cap B)}$$

- Dimostrazione grafica della Legge di De Morgan

CALCOLO COMBINATORIO E PROBABILITÀ

- Disposizioni, di ***n oggetti*** in ***gruppi di k***, con ripetizioni
  - Conta l'ordine, possibili le ripetizioni
    - Gli n oggetti possono tutti essere usati a ognuna delle k estrazioni

$$n^k$$

- Disposizioni, di ***n oggetti*** in ***gruppi di k***, semplici senza ripetizioni
  - Conta l'ordine, non si possono avere ripetizioni
    - Alla prima delle k estrazioni ho n possibilità, alla seconda n-1, alla k-esima n-k+1
    - NOTA:
      - Disposizioni, di ***n oggetti*** in ***gruppi di n,*** (*permutazioni*) ***n!***

$$\frac{n!}{(n-k)!}$$

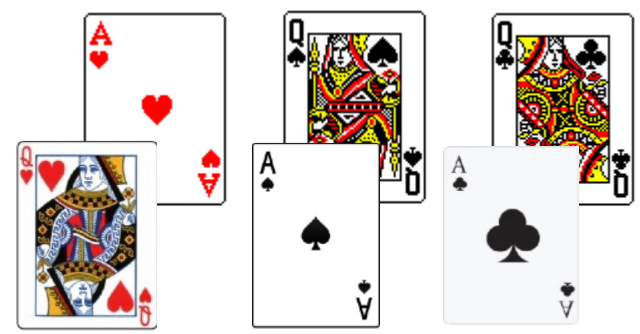
- Combinazioni, di ***n oggetti*** in ***gruppi di k***,
  - Non conta l'ordine e non si possono avere ripetizione; il numero di combinazioni = numero di disposizioni / permutazioni di k oggetti

$$\frac{n!}{k!\,(n-k)!}$$

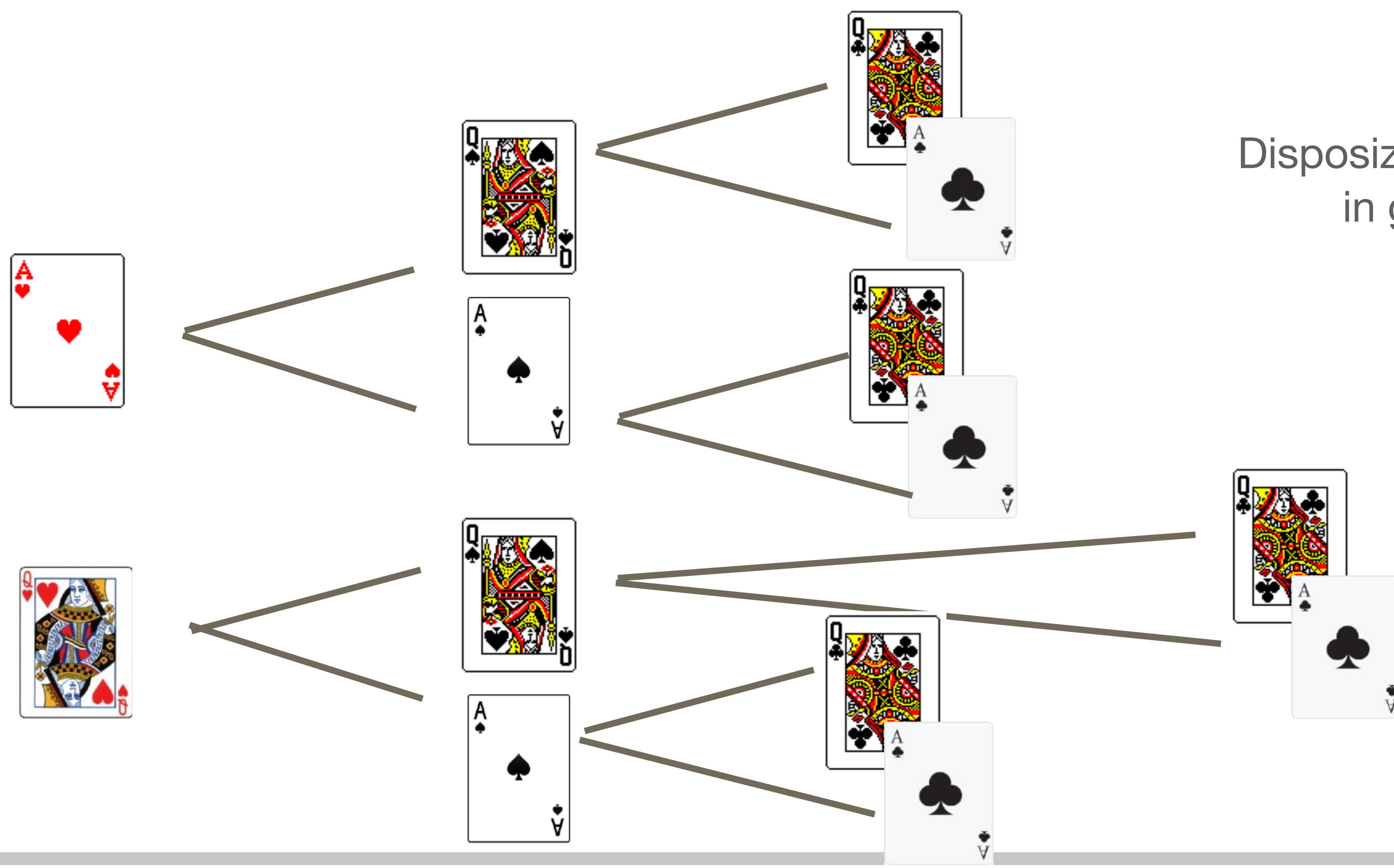
CALCOLO COMBINATORIO E PROBABILITÀ

- ***Disposizioni***

Ho a disposizione 2 carte di cuori, due di picche e 2 di fiori



Ne scelgo una di ogni seme, quante possibilità ho ?
=> ***Disposizioni***



Disposizioni di **n=2 (2 carte /seme)** in gruppi di **k=3 (3 semi)**

$2 \times 2 \times 2 = 2^3$

CALCOLO COMBINATORIO E PROBABILITÀ

- ***Disposizioni***

Ho a disposizione 10 cifre, da 0 a 9

Quanti numeri di 4 cifre posso formare
=> ***Disposizioni***
Le disposizioni sono ammesse ovviamente

$$10 \times 10 \times 10 \times 10 = 10^4$$

0000
0001
0002
…
1027
1028
1029
1030
1031
…
9997
9998
9999

CALCOLO COMBINATORIO E PROBABILITÀ

- ***Disposizioni semplici senza ripetizioni***

Quante possibili terne ordinate si possono formare estraendo 3 numeri dal bussolotto del lotto?

1-2-3, 1-2-4, 1-2-5, …., 90-89-88

Per il primo numero posso scegliere tra 90, per il secondo solo tra 89 per il terzo tra 87

$$\boxed{\frac{n!}{(n-k)!}}$$

90*(90-1)*(90-2)
In generale ***n(n-1)(n-2)….(n-k+1)***

- ***Combinazioni***

Se per le terne non conta l'ordine di estrazione, *siccome ogni terna può comparire in 3! permutazioni,*
cioè 1-2-3, 1-3-2, 2-3-1, ecc sono equivalenti

90*(90-1)*(90-2) / 3!
In generale:
n(n-1)(n-2)….(n-k+1) / k!

$$\boxed{\frac{n!}{k!\,(n-k)!} = \binom{n}{k}}$$

Coefficienti bonomiali

Calcolo combinatorio e probabilità

- Giocando a Poker, ho in mano una coppia di donne e un asso.
  - E' più conveniente chiedere tre carte scartando l'asso (caso A) o chiederne solo due e tenersi l'asso (caso B) ?

Identifichiamo lo ***spazio campione***

32 carte nel mazzo, 5 carte mie -> 27 possibili valori delle carte a disposizione

Caso A

$$\frac{27!}{3!\,(27-3)!} = 27 \times 26 \times 25/(3 \times 2) = 2925$$

Caso B

$$\frac{27!}{2!\,(27-2)!} = 27 \times 26 / 2 = 351$$

CALCOLO COMBINATORIO E PROBABILITÀ

- Per rispondere alla domanda se è più "conveniente" tenersi l'asso o no, occorre introdurre il ***concetto di probabilità***.
  - Una scelta è più conveniente dell'altra se rende più probabile il successo.

Definizione di probabilità:

1. Probabilità matematica o a priori
   Si basa sul concetto di "eventi ugualmente probabili"
2. Probabilità frequentistica o a posteriori
   Necessità di definire un "n molto grande"
3. Probabilità soggettiva
   Non ha fondamento matematico

4. Definizione assiomatica di probabilità (dovuta a Kolmogorov)

   Sia S uno spazio campione finito, ad ogni evento A di S si associa un numero reale P(A), detto probabilità dell'evento A, che soddisfa i seguenti assiomi:

   1. $0 \leq P(A) \leq 1$
   2. $P(S)=1$
   3. Se A e B sono eventi mutuamente esclusivi (cioè $A \cap B = 0$) allora $P(A \cup B) = P(A) + P(B)$

CALCOLO COMBINATORIO E PROBABILITÀ

Proprietà ricavabili dalla definizione assiomatica:

1. Se **A** e **B** sono due eventi **qualunque** in S allora (**regola additiva**)
   $P(A \cup B)=P(A)+P(B)-P(A \cap B)$

2. Se **A** è un evento di S allora $P(\overline{A}) = 1 - P(A)$

Probabilità condizionata definizione:

Siano A e B due eventi di S con $P(A) \neq 0$. La probabilità dell'evento B nell'ipotesi che si sia già verificato l'evento A è detta probabilità condizionata ed è definita come:

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

Da cui segue

Prob condizionata Prob condizionata

$$\boxed{P(A)P(B \mid A) = P(A \cap B) = P(B \cap A) = P(B)P(A \mid B)}$$

Proprietà commutativa

CALCOLO COMBINATORIO E PROBABILITÀ

- ***Esempio***: Supponiamo di voler determinare la probabilità che il numero ottenuto nel lancio di due dadi (somma) sia 7 sapendo che sul primo dado è uscito il numero 3.

- Il numero di combinazioni con il primo dado uguale a 3 e' 6
  - (3,1), (3,2), (3,3), (3,4), (3,5), (3,6)
- Sola la combinazione (3,4) mi da 7, quindi la probabilità che esca 7 dato il primo dado 3 è ***P(B|A) = 1/6***

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

- Ritroviamo questo risultato usando la relazione
- La probabilità che sia A={3 sul primo dado} che B={4 sul secondo dado} si verifichino nel lancio dei due dadi (probabilita' A e B) è data da 1 sola combinazione su 36,
  - 1/36
- La probabilità, invece, che il primo dado sia uguale a 3 è: P(A) = 1/6

- ***P(B|A) =(1/36) / (1/6) = 1/6***

CALCOLO COMBINATORIO E PROBABILITÀ

Due eventi si dicono indipendenti se:

$$P(B|A)=P(B)$$

Esempio: Probabilità che sul secondo dado esca 4 sapendo che sul primo dado è uscito 3

E quindi anche

$$P(A|B)=P(A)$$

Nel caso di eventi indipendenti si deduce che

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

Infatti:

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

ma

$$P(B|A)=P(B) \quad \text{e} \quad P(A|B)=P(A)$$

per cui

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

CALCOLO COMBINATORIO E PROBABILITÀ

- Giocando a Poker classico, ho in mano una coppia di donne e un asso.
  - E' più conveniente chiedere tre carte scartando l'asso o chiederne solo due e tenersi l'asso?
- Non sempre in un gioco d'azzardo è univocamente identificabile la soluzione "più conveniente". ***Consideriamo la possibilità di ottenere un Full o un Poker***.

Caso B

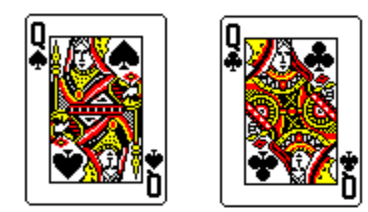
- Identifichiamo gli eventi favorevoli:
  - Eventi favorevoli un asso e una donna, due assi o due donne. Delle 27 carte nel mazzo solo **5 carte danno origine a eventi favorevoli** (perche' nel mazzo restano 3 assi e due donne)
  - Voglio valutare il numero di combinazioni favorevoli tra tutte le combinazioni possibili, cioe' in quanti modi posso ottenere combinazioni di due carte da un insieme di 5.
    - Eventi favorevoli: numero di combinazioni di 2 carte tra 5 = 10 = 5! / [(5-2)! 2!]
    - Probabilita' di poker o full, nel caso B, = 10/***351*** = 2.8%

351 = numero di elementi dello spazio campione

CALCOLO COMBINATORIO E PROBABILITÀ

- Giocando a Poker classico, ho in mano una coppia di donne e un asso.
  - E' più conveniente chiedere tre carte scartando l'asso o chiederne solo due e tenersi l'asso?
- Non sempre in un gioco d'azzardo è univocamente identificabile la soluzione "più conveniente". ***Consideriamo la possibilità di ottenere un Full o un Poker***.

Caso A



- Identifichiamo gli eventi favorevoli:
  - Eventi favorevoli
    - due donne + una carta X (diversa da donna): **POKER**
      - -> **25 combinazioni** (25 carte che possono accompagnare le due donne rimanenti)
    - tre carte (diverse da donna) uguali: **FULL (non di DONNA)**
      - 7 carte per seme (8 meno donna); di queste 3 (diverse tra loro) in mano; 4x4[combinazioni di 3 tra 4] + 3x1 = **19**
    - una (delle due rimanenti) donna + due carte X (diverse da donna) uguali: **FULL di DONNA**
      - -> 2(donne rimanenti) x { 4 x 12[combinazioni di 2 tra 4] + 3 x 3[combinazioni di 2 tra 3] } = 2 x 57 **= 114**
  - Eventi favorevoli: 25+19+114; Probabilita' di poker o full, nel caso A, = 158/***2925*** = 5.4%

Verificate

CALCOLO COMBINATORIO E PROBABILITÀ

- Esempio:
- Supponiamo di avere a che fare con una malattia rara che colpisce la popolazione con una frequenza di 2 malati ogni 100000 abitanti.
- P(malato)=2/100000=2 x 10^{-5}

NOTA: probabilità nota a priori

- Supponiamo di disporre di un test con le seguenti caratteristiche:
- ***Sensibilità***: probabilità, cioè, di un risultato positivo se la persona è effettivamente malata
  - P(pos | malato) = 0.993

Calibrazione. Conosco quale è l'effetto data la causa.

- ***Specificità***: probabilità, cioè, di un risultato negativo se la persona è sana
  - P(neg | sano) = 0.9999

Il test è molto buono. La Sensibilità mi sta dicendo che identifica un malato se questo lo è effettivamente con una probabilità del 99.3 % mentre la specificità mi dice che dichiara una persona sana se questa lo è effettivamente con una probabilità del 99.99%

CALCOLO COMBINATORIO E PROBABILITÀ

- Supponiamo di voler valutare se sia utile o meno sottoporre a screening la popolazione solo in base ai risultati di questo test.

- Caso pratico: sottopongo a test 10^6 soggetti (screening), cosa mi aspetto si ottenga dal test?

$$P(pos \mid malato) = 0.993 \quad \text{Sensibilità}$$

$$P(neg \mid sano) = 0.9999 \quad \text{Specificità}$$

- **Uso una simulazione per verificare le mie previsioni.**
- Prendo 10^6 soggetti e casualmente assumo che siano malati o sani in base alle conoscenze a priori. In base alla sensibilità e specificità del test simulo la risposta del test per ogni soggetto.
- ***Valuto quanti soggetti positivi al test sono effettivamente malati***
  - ***P (malato | pos)***

Teorema di Bayes

Sia A un evento di S con P(A)>0 e $\{B_1,..,B_n\}$ una famiglia di eventi dello spazio campione S mutuamente esclusivi ed esaustivi, allora

$$P(B_k \mid A) = \frac{P(A \mid B_k)P(B_k)}{P(A)}$$

$B_i \cap B_j = 0 \quad i \neq j$ — Mutuamente esclusivi

$B_1 \cup B_2 \cup \ldots \cup B_n = S$ — Esaustivi

Il teorema di Bayes è diretta conseguenza della definizione di probabilità condizionata

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

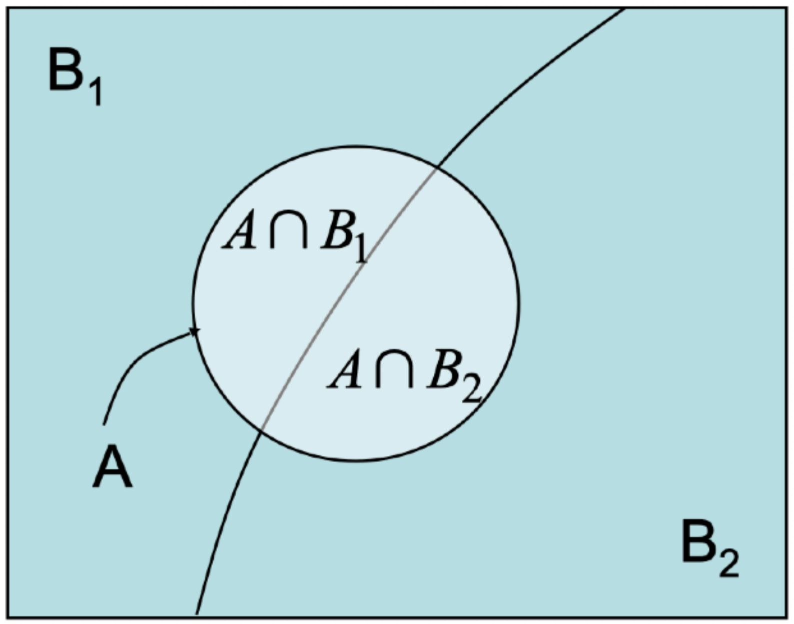
$$P(A)P(B \mid A) = P(A \cap B) = P(B \cap A) = P(B)P(A \mid B)$$

CALCOLO COMBINATORIO E PROBABILITÀ

Dividere lo spazio campione in n eventi mutuamente esclusivi e esaustivi equivale a dividerlo in n classi.

Es.: Classi== malati di cuore e non; Gruppi di massa in uno spettro di Raggi Cosmici, etc.

Il teorema di Bayes risponde alla domanda:
Supponiamo di aver ottenuto l'evento A qual è la probabilità che il mio elemento appartenga alla classe k?



B_1 e B_2 esclusivi ed esaustivi

$$P(A) = P(A \cap B_1) + P(A \cap B_2)$$

Ma per la definizione di probabilità condizionata

$$P(A \cap B_1) = P(A \mid B_1)P(B_1)$$

Da cui si dimostra il

Teorema della probabilità totale
Sia A un evento di S e $\{B_1, \ldots\, B_n\}$ una famiglia di eventi di S mutuamente esclusivi e esaustivi

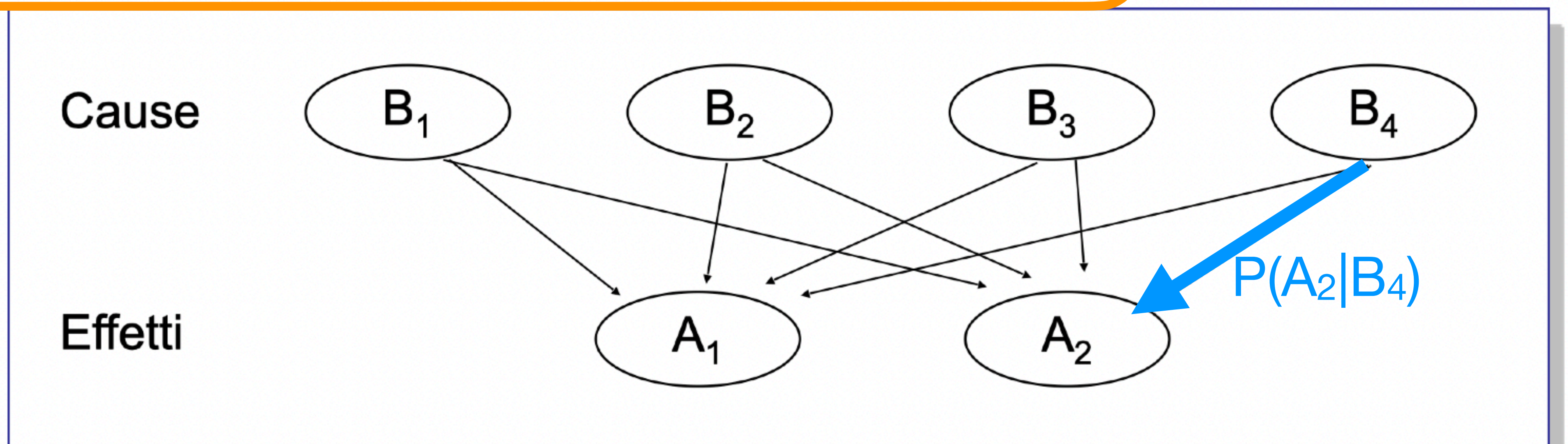
$$P(A) = P(A \mid B_1)P(B_1) + \ldots + P(A \mid B_n)P(B_n)$$

Usando il teorema della probabilità totale, il teorema di Bayes si può riscrivere come:

$$P(B_k \mid A) = \frac{P(A \mid B_k)P(B_k)}{\sum_{k=1}^{n} P(A \mid B_k)P(B_k)}$$

CALCOLO COMBINATORIO E PROBABILITÀ

Analizziamo una misura in termini di causa ed effetto



Lo spazio degli eventi è separato in n classi mutuamente esclusive ed esaustive che rappresentano tutte le possibili cause (B_1, …, B_4 nella figura).

Ogni causa può dare origine con probabilità diversa ad alcuni effetti (A_1,A_2 nella figura).

La probabilità che l'effetto A_1 sia dovuto alla causa B_2 è normalmente noto e rappresenta la probabilità che si verifichi l'effetto A_1 quando si è verificata la causa B_2.

$P(A_1|B_2)$

In una misura tipicamente vedo (misuro) l'effetto (A_1) e mi chiedo a quale possibile causa sia dovuto.

$P(B_k|A_1)$.

Il teorema di Bayes mi aiuta in questo.

$$P(B_k\,|\,A_i) = \frac{P(A_i\,|\,B_k)\;P(B_k)}{\sum_j P(A_i\,|\,B_j)\;P(B_j)}$$

CALCOLO COMBINATORIO E PROBABILITÀ

$$P(B_k|A_i) = \frac{P(A_i|B_k)\ P(B_k)}{\sum_j P(A_i|B_j)\ P(B_j)}$$

Notiamo come sia la specificità che la sensibilità del test sono due esempi di probabilità condizionata.

Misura

Ora supponiamo di voler applicare il test ad un campione di persone a caso la domanda che ci poniamo è:
"*quanti di coloro che sono risultati positivi al test sono effettivamente malati e quanti invece sono sani?* "
Questo equivale ad identificare la probabilità (condizionata) che un soggetto positivo al test sia effettivamente malato

Questa è conoscenza a priori

$$P(malato\,|\,pos) = \frac{P(pos\,|\,malato)P(malato)}{P(pos\,|\,malato)P(malato) + P(pos\,|\,sano)P(sano)}$$

CALCOLO COMBINATORIO E PROBABILITÀ

$$\boxed{P(malato\,|\,pos) = \frac{P(pos\,|\,malato)P(malato)}{P(pos\,|\,malato)P(malato) + P(pos\,|\,sano)P(sano)}}$$

$$P(pos\,|\,malato) = 0.993$$ Sensibilità del test

$$P(malato) = 2\cdot 10^{-5}$$ Conoscenza a priori

$$P(pos\,|\,sano) = 1 - P(neg\,|\,sano) = 1 - 0.9999 = 1\cdot 10^{-4}$$ Da specificità

$$P(sano) = 1 - P(malato) = 0.99998$$ Conoscenza a priori

Da cui

$$P(malato\,|\,pos) = 0.1656$$

Per cui di coloro che risultano positivi al test solo poco più del 16% sono effettivamente malati i rimanenti (84%) sono sani ma erroneamente indicati come malati (falsi positivi)

CALCOLO COMBINATORIO E PROBABILITÀ

Quant'è invece la frazione di coloro che risultano negativi al test e sono effettivamente sani
Questo equivale a calcolare la P(sano|neg)

$$P(sano \mid neg) = \frac{P(neg \mid sano)P(sano)}{P(neg \mid sano)P(sano) + P(neg \mid malato)P(malato)}$$

Con un calcolo analogo al precedente si trova che

$$P(sano \mid neg) = 0.9999999$$

Quindi se il risultato del test è negativo, la probabilità di essere sani è estremamente elevata o, inversamente, la probabilità che un malato risulti negativo al test (falso negativo) è estremamente bassa.

FINE