

PROBABILITÀ e STATISTICA


APPUNTI DEL CORSO DEL PROF. STEFANO ZAPPERI
A CURA DI MATTEO TAJANA

2020



UNIVERSITÀ
DEGLI STUDI
DI MILANO

Pagina lasciata intenzionalmente vuota.

Documento redatto orgogliosamente con  in L^AT_EX.

Indice


1	Introduzione alla probabilità	1
1.1	Probabilità classica e frequentista	1
1.2	Probabilità Bayesiana	2
1.3	Variabili casuali	4
1.3.1	Variabili discrete	4
1.3.2	Variabili continue	5
1.3.3	Indipendenza	6
2	Distribuzioni di probabilità univariate	7
2.1	La funzione caratteristica	8
2.2	Trasformazione di variabili	9
2.3	Calcolo combinatorio	10
2.3.1	Distribuzione binomiale	12
2.3.2	Distribuzione multinomiale	13
2.4	Distribuzione di Poisson	14
2.5	Distribuzione Gaussiana	17
2.6	Distribuzione del χ^2	18
2.7	Distribuzione Gamma	18
2.8	Distribuzione Beta	19
2.9	Distribuzione di Cauchy	20
2.10	Distribuzione di Lévy	20
2.11	Numeri di occupazione	21
2.11.1	Bosoni	21
2.11.2	Fermioni	22
2.12	Distribuzioni geometrica e ipergeometrica	22
3	Il teorema del limite centrale	25
4	Distribuzioni di probabilità multivariate	27
5	Distribuzioni estremali	31
5.1	Massimi	31
5.2	Minimi	31
5.3	Teorema di Fisher-Tippett-Gnedenko	32
5.4	I boxplots	33
6	Inferenza statistica	34
6.1	Minimi quadrati	37
6.2	Propagazione degli errori	37
6.3	Intervalli di confidenza	39
6.4	Rumore Barkhausen	40
7	Il principio di massima entropia	41
8	Test statistici	44
8.1	z-statistics	46
8.2	t-test (di Student)	46
8.3	Test del χ^2	46
8.4	Test di Kolmogorov-Smirnov	47
9	Principal component analysis	47
10	Networks	49


PROBABILITY COMPARISONS

0.01%	YOU GUESS THE LAST FOUR DIGITS OF SOMEONE'S SOCIAL SECURITY NUMBER ON THE FIRST TRY	39%	LEBRON JAMES GETS TWO FREE THROWS BUT MISSES ONE
0.1%	THREE RANDOMLY-CHOSEN PEOPLE ARE ALL LEFT-HANDED	40%	A RANDOM SCRABBLE TILE IS A LETTER IN "STEPH CURRY"
0.2%	YOU DRAW 2 RANDOM SCRABBLE TILES AND GET M AND M YOU DRAW 3 RANDOM M&M'S AND THEY'RE ALL RED	46%	THERE'S A MAGNITUDE 7 QUAKE IN LA WITHIN 30 YEARS
0.3%	YOU GUESS SOMEONE'S BIRTHDAY IN ONE TRY	48%	MILWAUKEE HAS A WHITE CHRISTMAS A RANDOM SCRABBLE TILE IS A LETTER IN "CARLY RAE JEPSEN"
0.5%	AN NBA TEAM DOWN BY 30 AT HALFTIME WINS YOU GET 4 M&M'S AND THEY'RE ALL BROWN OR YELLOW	50%	YOU GET HEADS IN A COIN TOSS
1%	STEPH CURRY GETS TWO FREE THROWS AND MISSES BOTH LEBRON JAMES GUESSES YOUR BIRTHDAY, IF EACH GUESS COSTS ONE FREE THROW AND HE LOSES IF HE MISSES	53%	SALT LAKE CITY HAS A WHITE CHRISTMAS
1.5%	YOU GET TWO M&M'S AND THEY'RE BOTH RED YOU SHARE A BIRTHDAY WITH A BACKSTREET BOY	54%	LEBRON JAMES GETS TWO FREE THROWS AND MAKES BOTH
2%	YOU GUESS SOMEONE'S CARD ON THE FIRST TRY	58%	A RANDOM SCRABBLE TILE IS A LETTER IN "NATE SILVER"
3%	YOU GUESS 5 COIN TOSSES AND GET THEM ALL RIGHT STEPH CURRY WINS THAT BIRTHDAY FREE THROW GAME	60%	YOU GET TWO M&M'S AND NEITHER IS BLUE
4%	YOU SWEEP A 3-GAME ROCK PAPER SCISSORS SERIES PORTLAND, OREGON HAS A WHITE CHRISTMAS YOU SHARE A BIRTHDAY WITH TWO US SENATORS	65%	BURLINGTON, VERMONT HAS A WHITE CHRISTMAS
5%	AN NBA TEAM DOWN 20 AT HALFTIME WINS YOU ROLL A NATURAL 20	66%	A RANDOMLY CHOSEN MOVIE FROM THE MAIN LORD OF THE RINGS TRILOGY HAS "OF THE" IN THE TITLE TWICE
6%	YOU CORRECTLY GUESS SOMEONE'S CARD GIVEN 3 TRIES	67%	YOU ROLL AT LEAST A 3 WITH A D6
7%	LEBRON JAMES GETS TWO FREE THROWS AND MISSES BOTH	71%	A RANDOM SCRABBLE TILE BEATS A RANDOM DICE ROLL
8%	YOU CORRECTLY GUESS SOMEONE'S CARD GIVEN 4 TRIES	73%	LEBRON JAMES MAKES A FREE THROW
9%	STEPH CURRY MISSES A FREE THROW	75%	YOU DROP TWO PLAIN M&M'S AND ONE OF THEM LANDS WITH THE "M" UP SO IT'S CLEAR THEY'RE NOT SKITTLES
10%	YOU DRAW 5 CARDS AND GET THE ACE OF SPADES THERE'S A MAGNITUDE 8+ EARTHQUAKE IN THE NEXT MONTH	76%	YOU GET TWO M&M'S AND NEITHER IS RED
11%	YOU SWEEP A 2-GAME ROCK PAPER SCISSORS SERIES A RANDOMLY-CHOSEN AMERICAN LIVES IN CALIFORNIA	77%	YOU GET AN M&M AND IT'S NOT BLUE
12%	YOU CORRECTLY GUESS SOMEONE'S CARD GIVEN 6 TRIES YOU SHARE A BIRTHDAY WITH A US PRESIDENT	78%	AN NBA TEAM WINS WHEN THEY'RE UP 10 AT HALFTIME
13%	A D6 BEATS A D20 AN NBA TEAM DOWN 10 GOING INTO THE 4TH QUARTER WINS YOU PULL ONE M&M FROM A BAG AND IT'S RED	79%	ST. LOUIS DOESN'T HAVE A WHITE CHRISTMAS
14%	A RANDOMLY DRAWN SCRABBLE TILE BEATS A D6 DIE ROLL	81%	TWO RANDOM PEOPLE ARE BOTH RIGHT-HANDED
15%	YOU ROLL A D20 AND GET AT LEAST 18	83%	STEPH CURRY GETS TWO FREE THROWS AND MAKES BOTH
16%	STEPH CURRY GETS TWO FREE THROWS BUT ONLY MAKES ONE	85%	YOU ROLL A D20 AND GET AT LEAST 4
17%	YOU ROLL A D6 DIE AND GET A 6	87%	AN NBA TEAM UP BY 10 GOING INTO THE 4TH QUARTER WINS SOMEONE FAILS TO GUESS YOUR CARD GIVEN 7 TRIES
18%	A D6 BEATS OR TIES A D20	88%	A RANDOMLY CHOSEN AMERICAN LIVES OUTSIDE CALIFORNIA
19%	AT LEAST ONE PERSON IN A RANDOM PAIR IS LEFT-HANDED	89%	YOU ROLL A 3 OR HIGHER GIVEN TWO TRIES
20%	YOU GET A DOZEN M&M'S AND NONE OF THEM ARE BROWN	90%	SOMEONE FAILS TO GUESS YOUR CARD GIVEN 5 TRIES
21%	ST. LOUIS HAS A WHITE CHRISTMAS	91%	YOU INCORRECTLY GUESS THAT SOMEONE WAS BORN IN AUGUST STEPH CURRY MAKES A FREE THROW
22%	AN NBA TEAM WINS WHEN THEY'RE DOWN 10 AT HALFTIME	92%	YOU GUESS SOMEONE'S BIRTH MONTH AT RANDOM AND ARE WRONG
23%	YOU GET AN M&M AND IT'S BLUE YOU SHARE A BIRTHDAY WITH A US SENATOR	93%	LEBRON JAMES MAKES A FREE THROW GIVEN TWO TRIES
24%	YOU CORRECTLY GUESS THAT SOMEONE WAS BORN IN THE WINTER	94%	SOMEONE FAILS TO GUESS YOUR CARD GIVEN 3 TRIES
25%	YOU CORRECTLY GUESS THAT SOMEONE WAS BORN IN THE FALL YOU ROLL TWO PLAIN M&M'S AND GET M AND M	95%	AN NBA TEAM WINS WHEN THEY'RE UP 20 AT HALFTIME
26%	YOU CORRECTLY GUESS SOMEONE WAS BORN IN THE SUMMER	96%	SOMEONE FAILS TO GUESS YOUR CARD GIVEN 2 TRIES
27%	LEBRON JAMES MISSES A FREE THROW	97%	YOU TRY TO GUESS 5 COIN TOSSES AND FAIL
32%	PITTSBURGH HAS A WHITE CHRISTMAS A RANDOMLY CHOSEN STAR WARS MOVIE (EPISODES I-IX) HAS "OF THE" IN THE TITLE	98%	YOU INCORRECTLY GUESS SOMEONE'S BIRTHDAY IS THIS WEEK
33%	YOU WIN THE MONTY HALL SPORTS CAR BY PICKING A DOOR AND REFUSING TO SWITCH YOU WIN ROCK PAPER SCISSORS BY PICKING RANDOMLY	98.5%	AN NBA TEAM UP 15 POINTS WITH 8 MINUTES LEFT WINS
34%	YOU DRAW 5 CARDS AND GET AN ACE	99%	STEPH CURRY MAKES A FREE THROW GIVEN TWO TRIES
35%	A RANDOM SCRABBLE TILE IS ONE OF THE LETTERS IN "RANDOM"	99.5%	AN NBA TEAM THAT'S UP BY 30 POINTS AT HALFTIME WINS
		99.7%	YOU GUESS SOMEONE'S BIRTHDAY AT RANDOM AND ARE WRONG
		99.8%	THERE'S NOT A MAGNITUDE 8 QUAKE IN CALIFORNIA NEXT YEAR
		99.9%	A RANDOM GROUP OF THREE PEOPLE CONTAINS A RIGHT-HANDER
		99.99%	YOU INCORRECTLY GUESS THE LAST FOUR DIGITS OF SOMEONE'S SOCIAL SECURITY NUMBER
		99.999999999999995%	YOU PICK UP A PHONE, DIAL A RANDOM 10-DIGIT NUMBER, AND SAY "HELLO BARACK OBAMA, THERE'S JUST BEEN A MAGNITUDE 8 EARTHQUAKE IN CALIFORNIA!" AND ARE WRONG
		0.00000001%	YOU ADD "HANG ON, THIS IS BIG—I'M GOING TO LOOP IN CARLY RAE JEPSEN," DIAL ANOTHER RANDOM 10-DIGIT NUMBER, AND SHE PICKS UP

SOURCES: XKCD.COM/2379/SOURCES

Prima volta che il corso viene erogato. È stato introdotto per far parte del percorso di “Fisica dei dati”, e si è pensato fosse utile avere un corso di partenza che fornisse il linguaggio di base per le analisi avanzate di dati. È un corso di base per il machine learning/deep learning. La fisica dei dati è un argomento che permea diversi settori: fisica medica, fisica sperimentale delle alte energie, ... può servire a vari scopi.

Le lezioni saranno di due tipi: ci saranno una parte teorica e poi una di “esercitazione” su  (jupyter-notebook) su cui vedremo esempi e piccoli esercizi. Essendo un corso di probabilità e statistica bisogna trattare entrambi gli elementi: la prima è la base teorica della statistica pratica. Bisogna creare degli strumenti statistici per analizzare dei problemi pratici. Ciò che vediamo è artefatto o è vero?

Il sito del corso è <http://labonline.ctu.unimi.it/course/view.php?id=170>, è più interattivo di Ariel. Le esercitazioni saranno su  all'url <https://github.com/SZapperi/CorsoProbabilitaStatistica>. Ci sono vari argomenti, che corrispondono a lezioni diverse. Parleremo della definizione di probabilità (e differenza frequentismo/Bayes), poi ci sono le distribuzioni univariate e le funzioni caratteristiche (cose abbastanza formali, siamo lontani dalla statistica). Poi avremo una parte sul calcolo combinatorio, come contare gli oggetti nei vari casi, partizioni e numeri di occupazione. La distribuzione di Poisson sarà successivamente approfondita assieme ai numeri di occupazione. Il Teorema del Limite Centrale verrà ricavato e generalizzato. Poi passeremo alle distribuzioni multivariate, con correlazioni tra le variabili. Verso la fine del corso affronteremo argomenti più “esotici”, con distribuzioni estremali (legge dei grandi numeri, fisica delle fratture), maximum likelihood, confidence levels e MaxEnt. Infine concluderemo con i test statistici per vedere se l'effetto mediano è vero o è dovuto al caso: se giochiamo a dadi e il nostro avversario tira due 6 uno di seguito all'altro, è un baro o è tutto regolare? Ci saranno poi degli esercizi di approfondimento a seconda del punto in cui saremo arrivati con il corso.

La statistica è una disciplina pratica, spiegarla in maniera puramente teorica è poco utile, bisogna sporcarsi le mani con i dati per capirla veramente.

1 Introduzione alla probabilità

1.1 Probabilità classica e frequentista

La definizione “classica” della probabilità è dovuta a Pascal e Fermat, ed è quella secondo cui *la probabilità è data dal numero di eventi favorevoli normalizzata su quelli totali*. La probabilità di avere un “6” al tiro di dado è data da $P = g/N$, dove g è il numero di uscite di 6 mentre N è il numero totale di tiri. Questa definizione ha una serie di proprietà. In particolare: (i) $0 \leq P(A) \leq 1$, (ii) $P(\text{totale}) = 1$ e (iii) $P(\text{impossibile}) = 0$. Inoltre $P(A \cup B) = \frac{1}{N}(n_A + n_B - n_{A \cap B}) = P(A) + P(B) - P(A \cap B)$. Per eventi esclusivi sappiamo che $A \cap B = \emptyset$ e che $P(A \cup B) = P(A) + P(B)$. Invece per eventi complementari vale che $A \cap \bar{A} = \emptyset$ e $P(A) + P(\bar{A}) = 1$.

Questa definizione ha una serie di problemi. Consideriamo per esempio due dadi. Ci chiediamo quale sia la probabilità che la somma del tiro dei due dadi sia pari a 7, cioè $P(S = 7)$. La risposta è ovviamente $1/6$, ed emerge qui la tematica dell'indistinguibilità (classica, non quantistica). La definizione data prima si mostra quindi ambigua, perché intuitivamente darebbe $1/11$ —la somma di due dadi a sei facce infatti produce 11 numeri (i naturali tra 2 e 12, estremi inclusi), e l'evento $P(S = 7)$ è un singolo evento nello spazio degli 11 possibili. Non è una definizione assoluta, dipende dal contesto e da dove la applichiamo.

A questo punto ci chiediamo quale possa essere una definizione “migliore”. Possiamo pensare ad una definizione classica data da Bernoulli e Laplace: *la probabilità è il rapporto tra eventi favorevoli ed eventi possibili SE questi sono equiprobabili*. Non tutti gli eventi avvengono lo stesso numero di volte: l'evento somma = 2 avviene una volta, mentre quello somma = 3 avviene due volte—vedi Fig. 1.

La definizione più comune di probabilità, quella statistica, è la cosiddetta definizione statistica/frequentista, cioè la probabilità come limite della frequenza.

$$P = \lim_{N \rightarrow \infty} \frac{n}{N}.$$

Possiamo sempre usare questa definizione? In realtà ci sono problemi anche per questa definizione. Per esempio, qual è la probabilità che domani piova? La domanda è ben posta, ma questa definizione non può essere applicata in questo caso. Non ha troppo senso andare a vedere nei registri meteorologici vecchi quanti giorni ha piovuto per fare una statistica contando i giorni di sole o pioggia. Arriviamo quindi alla probabilità soggettiva, perché non sempre è oggettiva. E si arriva quindi alla definizione Bayesiana della probabilità.



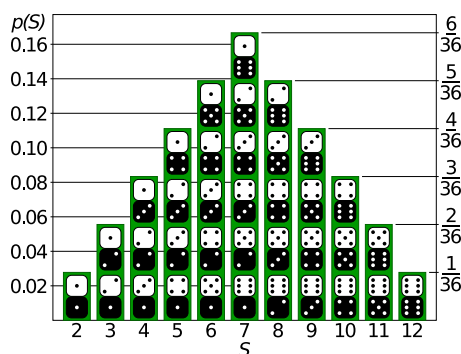


Figura 1: Funzione della probabilità di massa della somma di due dadi.

1.2 Probabilità Bayesiana

L'approccio Bayesiano si usa quando abbiamo un dato relativo: ad esempio, la pioggia dipende dalle condizioni meteorologiche recenti (ieri/oggi...).

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

Dobbiamo tener conto tutto ciò che ha preceduto l'evento, il passato. Ad esempio per la moneta, in termini frequentisti tiriamo tot volte e vediamo quante volte esce testa o croce. Nell'approccio Bayesiano invece abbiamo $P(s = t|I = I_0)$ dove I_0 è il bias che la moneta (non) sia truccata e quindi $P = \frac{1}{2}$. Invece sia I_1 il bias che la moneta sia truccata. Allora in questo caso $P > \frac{1}{2}$. Nel caso della moneta non è particolarmente illuminante, ma ci sono casi in cui la probabilità bayesiana sia più efficace di quella frequentista, come ad esempio per il meteo. La definizione "classica" porta a dei problemi in alcuni casi → definizione "frequentista" più adeguata. Infine abbiamo completato con la definizione "Bayesiana" della probabilità, in cui non esiste una probabilità assoluta: esistono solo probabilità condizionate, dipendenti dalle informazioni che abbiamo.

Continuiamo con la probabilità condizionata, dipendente dall'informazione I che abbiamo: $P(S|I)$. Quali sono le regole della probabilità condizionata?

- (i) regola di somma: $P(A \vee B|I) = P(A|I) + P(B|I) - P(A \wedge B|I)$;
- (ii) regola del prodotto, per cui $P(A \wedge B|I) = P(A|B, I)P(B|I)$. Questa cosa può essere scritta anche all'inverso ovviamente;
- (iii) regola di normalizzazione è $P(A|I) + P(\bar{A}|I) = 1$;
- (iv) marginalizzazione $P(A|I) = P(A \wedge B|I) + P(A \wedge \bar{B}|I)$.

La regola del prodotto è alla base del **Teorema di Bayes**:¹

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

E quindi

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|\bar{A})P(\bar{A})}.$$

Di fatto quando bisogna fare dei calcoli si usa solitamente la "seconda versione" del teorema di Bayes, in cui si scompone la probabilità totale di osservare l'evidenza di un fatto. In altre parole,

$$\text{Posterior} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Evidence}}.$$

Quello che esprime il teorema è come calcolare quanto spesso A sia vero tra tutti i casi in cui B lo è.

Esercizio: data una popolazione di studenti, vogliamo stimare la frazione di studenti che sono donne. Sappiamo (1) la frazione di studenti nella popolazione, 2%. Sappiamo anche che (2) la frazione femminile nella popolazione totale è del 50%, e che (3) nella popolazione femminile le studentesse sono l'1.8%.

Per applicare il teorema di Bayes dobbiamo individuare le varie probabilità condizionate. Noi cerchiamo la probabilità $P(\text{essere studente}|\text{essere donna})$. Di conseguenza troviamo che $P(A) = 2\%$, e $P(B) = 50\%$ e $P(A|B) = 1.8\%$, e otteniamo quindi $P(B|A) = 0.45\%$.



¹Vedi video di 3Blue1Brown : <https://www.youtube.com/watch?v=HZGCoVF3YvM> e di Veritasium : <https://www.youtube.com/watch?v=R13BD8qKeTg>.

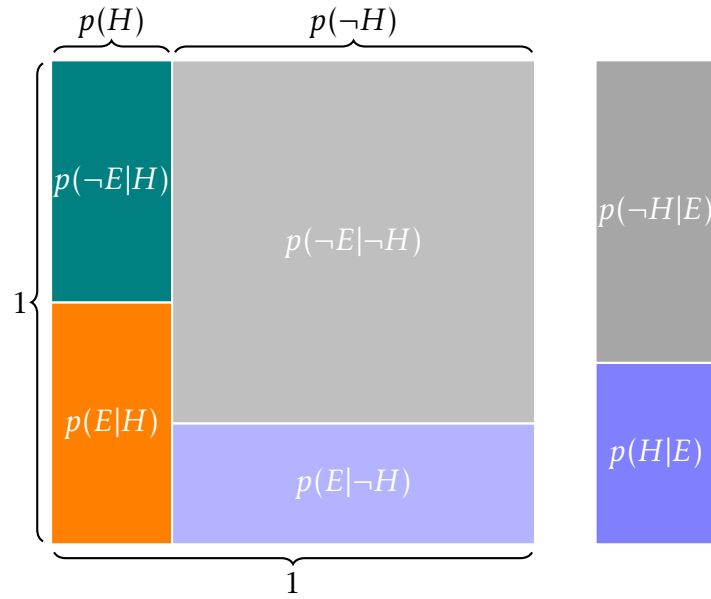


Figura 2: Rappresentazione geometrica del Teorema di Bayes sul quadrato unitario 1×1 . Osserviamo come $P(E|H)$ sia dato dalla *proporzione* di $P(E)$ rispetto a $P(H)$.

La teoria della probabilità trova sicuramente (massima?) espressione nel gioco d'azzardo. Più il gioco d'azzardo è complicato e più è interessante studiarla. Un esempio semplice è il gioco delle tre carte: l'obiettivo è trovare la carta vincente. Può diventare un gioco puramente probabilistico se le tre carte sono veramente mischiate. Miglioriamo il gioco, supponendo di scoprire una carta che il mazziere sa non essere quella giusta. La domanda è questa: se ci viene data possibilità di cambiare carta, conviene cambiare o è indifferente? Anche se sembra contro intuitivo, conviene cambiare (paradosso di Monty Hall). Un modo per comprendere questo fatto è supporre che ci siano 100 carte. Ne scegliamo una, e il mazziere ne scopre 98. Ovviamente in questo caso conviene cambiare, perché le probabilità sono 99% e 1% *rispetto alla scelta iniziale*. La Fig. 3 rappresenta la situazione con 20 porte anziché 100 carte. Altrimenti ci si può fare la tabella, la matrice della scelta. Si può dimostrare anche con la probabilità Bayesiana! Sia S il tenere la carta, e G la vittoria. Vogliamo vedere quanto è $P(G|S)$. La marginalizzazione ci dice che $\sigma = 1$ se la prima scelta è giusta, e $\sigma = 0$ se invece è sbagliata. Quindi

$$P(G|S) = \sum_{\sigma=0}^1 P(G|S, \sigma)P(\sigma|S) = P(G|S, 0)P(S|0) + P(G|S, 1)P(S|1),$$

dove $P(0|S)$ è la probabilità che la prima scelta sia sbagliata e che non abbiamo cambiato carta. La probabilità di vittoria non cambiando carta è quindi pari a $P(S|1) = \frac{1}{3}$. Adesso invece cerchiamo il caso C cambio carta, e G vittoria. Analogamente a prima, calcoliamo ora

$$P(G|C) = P(G|C, 0)P(C|0) + P(G|C, 1)P(C|1).$$

Se la scelta era ok ma cambio la probabilità di vincere è ovviamente $P(C|1) = 0$, mentre se ho cambiato e la prima scelta era sbagliata e quindi vinco, $P(G|1, 0) = 1$ e quindi $P(C|0) = \frac{2}{3}$.

La probabilità non è qualcosa di intuitivo, forse "alla fine conviene contare".

Esercizio: Sia dato un test diagnostico con le seguenti caratteristiche:

- Sensibilità P_{SE} = probabilità di identificare correttamente un malato = 95%
- Specificità P_{SP} = probabilità di identificare correttamente un sano = 90%

Sia $f = 5\%$ la prevalenza di una certa malattia (percentuale della popolazione che mostra i sintomi). Qual è la probabilità di essere malato se il test è positivo?

Conviene come prima cosa fare una tabella che incroci le il numero di persone positive/negative a seconda dell'esito del test, data una popolazione di N individui. Sappiamo che i malati sono Nf moltiplicato per la sensibilità del test. Analogamente possiamo calcolare gli altri valori:



Test	Malato	Non Malato
+	NfP_{SE}	$N(1-f)(1-P_{SP})$
-	$Nf(1-P_{SE})$	NfP_{SP}

Dato un test positivo, vogliamo la probabilità che il l'individuo in considerazione sia veramente malato. Questo sarà dato dalla probabilità che il test sia positivo fP_{SE} diviso le altre due possibilità normalizziamo per la linea positiva—cioè bisogna dividere per la probabilità di essere positivo, data dalla combinazione di essere malato e positivo assieme a quella di essere sano e falso positivo. La soluzione è quindi

$$\frac{fP_{SE}}{fP_{SE} + (1-f)(1-P_{SP})} = \frac{0.05 \times 0.95}{0.05 \times 0.95 + 0.95 \times 0.1} \approx \frac{1}{3}.$$

Osserviamo quanto la probabilità di diagnosticare correttamente un vero positivo con un solo test sia bassa: solo due tamponi su tre danno un esito corretto! Questo è dovuto al fatto che il suddetto test ha una prevalenza della popolazione bassa (5%). Test sierologici con sensibilità bassa (95% non è altissima, i tamponi per il COVID-19 hanno una sensibilità maggiore del 99%) non hanno rilevanza sul singolo ma sulla popolazione (screening della popolazione, i falsi positivi sono pochi su una grande popolazione). Un'ultima osservazione interessante è la seguente: come cambia la probabilità se viene eseguito due volte (in maniera indipendente) il tampone sullo stesso individuo? Riscrivendo la formula di Bayes aggiornando le probabilità otteniamo:

$$\frac{P_{SE}P^*}{P_{SE}P^* + (1-P_{SP})(1-P^*)} = 0.95 \times \frac{\frac{1}{3}}{0.95 \times \frac{1}{3} + 0.1 \times \frac{2}{3}} \approx 80\%.$$

La nuova probabilità basata su due test positivi è molto più alta! Nuove osservazioni aggiornano le nostre prior probabilities $P(A)$. È difficile che due test indipendenti diano risultati diversi se la sensibilità è alta...

Should you switch or stay?

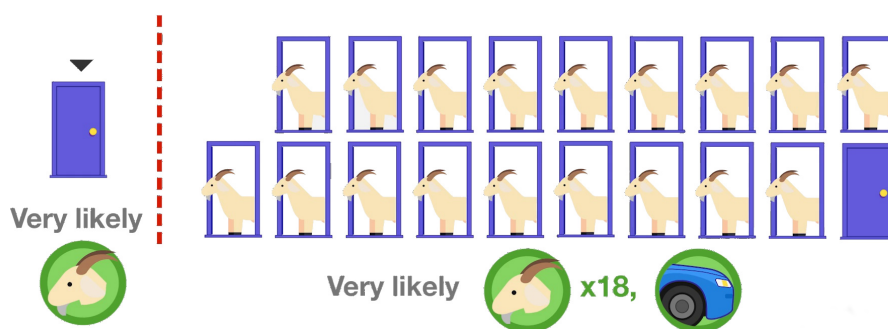


Figura 3: Esempio più grande del paradosso di Monty Hall con 20 porte, 19 capre e 1 macchina.

1.3 Variabili casuali

Vogliamo andare ad analizzare le proprietà delle variabili casuali, sia discrete che continue.

1.3.1 Variabili discrete

L'esempio classico è quello della moneta. Abbiamo $X_n = 1, \dots, N$ a cui associamo una probabilità P_n . Nel caso di una moneta $N = 2$ e se non è truccata $P = \frac{1}{2}$. Il valore atteso è definito da

$$\langle X \rangle = \sum_{n=1}^N X_n P_n.$$



Si dimostra che $\langle \alpha X + \beta Y \rangle = \alpha \langle X \rangle + \beta \langle Y \rangle$. Possiamo generalizzare la media con i momenti (medie di ordine superiore). Il momento k -esimo è definito come $m_k := \langle X^k \rangle$. I momenti centrali sono definiti come $\mu_k := \langle (x - \langle X \rangle)^k \rangle$ e ovviamente $\mu_1 = 0$ e $m_0 = 1$. Il più noto momento centrale è la varianza, indice di quanto i valori si disperdono rispetto alla media:

$$\sigma^2 = \mu_2 = \langle (x - \langle x \rangle)^2 \rangle = \langle x^2 \rangle - \langle x \rangle^2 = m_2 - m_1^2.$$

Infine definiamo la deviazione standard come la radice quadrata della varianza.

$$\text{STD} = \sigma = \sqrt{\mu_2} = \sqrt{\langle x^2 \rangle - \langle x \rangle^2}.$$

Questa è la misura standard delle incertezze stocastiche. Tutte queste proprietà sono valide anche per le variabili continue, che andiamo ora ad analizzare.

1.3.2 Variabili continue

In questo caso $x \in \mathbb{R}$, e non possiamo associare una probabilità ad una singola variabile \rightarrow distribuzione. Distribuzione cumulata: $F(t) := \text{Prob}(x < t)$, $t \in (-\infty, \infty)$, dove ovviamente poi $F(-\infty) = 0$ e $F(\infty) = 1$.

Altra distribuzione interessante è quella di sopravvivenza, definita come $S(t) := 1 - F(t) = \text{Prob}(x > t)$, è un po' come la "complementare" della cumulata. Se le andassimo a graficare otterremmo i grafici della Fig. 4.

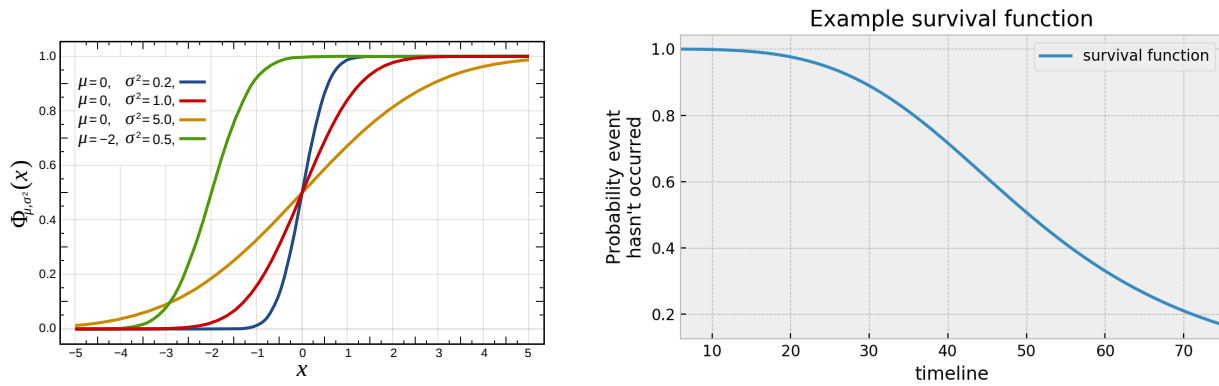


Figura 4: Cumulativa di diverse distribuzioni gaussiane ed esempio di una funzione di sopravvivenza.

Dalla distribuzione cumulata si può calcolare la funzione di distribuzione di probabilità (PDF, o pdm in inglese per probability density mass), definita dalla derivata della cumulata:

$$\rho(x) = dF/dx.$$

Devono valere $\rho(-\infty) = \rho(\infty) = 0$ e la normalizzazione $\int_{-\infty}^{\infty} dx \rho(x) = F(\infty) - F(-\infty) = 1$. Inoltre

$$\text{Prob}(x_1 \leq x \leq x_2) = F(x_2) - F(x_1) = \int_{x_1}^{x_2} dx \rho(x).$$

Anche per le variabili casuali continue valgono le definizioni date prima per media e momenti:

$$\langle x \rangle = \int_{-\infty}^{\infty} dx x \rho(x) = m.$$

Possiamo definire quindi la varianza

$$\sigma^2 = \int dx (x - m)^2 \rho(x).$$

Altro momento interessante è la skewness, che ci indica quanto è simmetrica la distribuzione:

$$\gamma_1 = \langle (x - m)^3 \rangle / \sigma^3.$$

Interessante è anche la curtosi, misura dell'ampiezza delle code della distribuzione, definita da

$$\beta_2 = \langle (x - m)^4 \rangle / \sigma^4.$$



per una gaussiana $\gamma_1 = \beta_2 = 0$. I momenti sono definiti da $m_k = \langle x^k \rangle$ e quelli centrali sono definiti allo stesso modo $\mu_k = \langle (x - m)^k \rangle$. Le medie sono definite da

$$\langle f(x) \rangle = \int dx f(x) \rho(x).$$

1.3.3 Indipendenza

Consideriamo due variabili casuali indipendenti definite da due PDF: $X \mapsto f(X)$, $Y \mapsto g(Y)$ —non ci interessa se continue o discrete; quello che ci interessa è la scorrelazione (indipendenza) delle variabili:

$$\langle XY \rangle = \langle X \rangle \langle Y \rangle.$$

Altra proprietà interessante ci viene data dalla varianza della somma: abbiamo anticipato che la media della somma è la somma delle medie. Per quanto riguarda la varianza, supponendo $X = X_1 + X_2$:

$$\sigma^2 = \langle [X - \langle X \rangle]^2 \rangle = \langle [(X_1 - m_1) + (X_2 - m_2)]^2 \rangle = \langle (X_1 - m_1)^2 \rangle + \langle (X_2 - m_2)^2 \rangle + 2\langle (X_1 - m_1)(X_2 - m_2) \rangle.$$

Adesso sviluppiamo i quadrati e il prodotto, ottenendo

$$\begin{aligned} \sigma^2 &= \langle X_1^2 \rangle + m_1^2 - 2\langle X_1 \rangle m_1 + \langle X_2^2 \rangle + m_2^2 - 2\langle X_2 \rangle m_2 + 2\langle X_1 X_2 - X_1 m_2 - X_2 m_1 + m_1 m_2 \rangle \\ &= \langle X_1^2 \rangle + m_1^2 - 2\langle X_1 \rangle m_1 + \langle X_2^2 \rangle + m_2^2 - 2\langle X_2 \rangle m_2 + 2\langle X_1 X_2 \rangle - 2\langle X_1 \rangle m_2 - 2\langle X_2 \rangle m_1 + 2m_1 m_2 \\ &= \langle X_1^2 \rangle + m_1^2 - 2\langle X_1 \rangle m_1 + \langle X_2^2 \rangle + m_2^2 - 2\langle X_2 \rangle m_2 + 2\langle X_1 \rangle \langle X_2 \rangle - 2\langle X_1 \rangle m_2 - 2\langle X_2 \rangle m_1 + 2m_1 m_2 \\ &\quad \text{ma valendo } \langle X_i \rangle \equiv m_i \text{ allora} \\ &= \langle X_1^2 \rangle + m_1^2 - 2m_1^2 + \langle X_2^2 \rangle + m_2^2 - 2m_2^2 + 2m_1 m_2 - 2m_1 m_2 - 2m_2 m_1 + 2m_1 m_2 \\ &= \langle X_1^2 \rangle - m_1^2 + \langle X_2^2 \rangle - m_2^2 \\ &\quad \text{ma essendo } m_i^2 \equiv \langle X_i \rangle^2, \text{ dalla definizione di STD } \sigma^2 = \langle X^2 \rangle - \langle X \rangle^2 \\ &\equiv \sigma_1^2 + \sigma_2^2. \end{aligned}$$

Consideriamo ora il caso più complesso di N variabili: $X = \sum_{i=1}^N X_i$. Allora $\sigma_N^2 = \sum_i \sigma_i^2$, e se $\sigma_i = \sigma \forall i$ allora $\sigma_N^2 = N\sigma^2 \Rightarrow$ “random walk”.

Nel caso di un insieme finito di N variabili random (media semplice) indicheremo la media in maniera differente rispetto alle parentesi angolari: $\bar{X} = \sum_i X_i / N$ e quindi $\langle \bar{X} \rangle = \sum_i \langle X_i \rangle / N = m = \langle X \rangle$. Qual è invece la migliore stima della varianza? O meglio, qual è la varianza della variabile \bar{X} ? Utilizzando il risultato di prima si ottiene

$$\sigma_{\bar{X}}^2 = \langle (\bar{X} - m)^2 \rangle = \frac{1}{N^2} \sigma_N^2 = \frac{1}{N} \sigma^2.$$

Possiamo riscriverlo come

$$\sigma_{\bar{X}} = \frac{1}{\sqrt{N}} \sigma,$$

dobbiamo dividere per la radice quadrata del numero totale di variabili \rightarrow errore sulla media ridotto quando prendiamo misure in un esperimento, più prendo variabili/misurazioni e più è precisa la mia osservazione, l'errore scala come l'inverso della radice del numero di campioni. Ricordiamoci sempre che tutto questo regge solo con l'assunzione di indipendenza delle variabili.

Sia $v_m^2 = \frac{1}{N} \sum_i (X_i - m)^2$. Consideriamo il caso in cui non si conosce m : la cosa migliore che si può fare è sostituire m con la media aritmetica. Allora $v^2 = \frac{1}{N} \sum_i (X_i - \bar{X})^2 = \frac{1}{N} \sum_i X_i^2 - \bar{X}^2$. Sappiamo anche che per ogni singola variabile

$$\langle X_i^2 \rangle = \sigma^2 + m^2, \quad \langle \bar{X}^2 \rangle = \sigma_{\bar{X}}^2 + \langle \bar{X} \rangle^2 \Rightarrow \langle \bar{X}^2 \rangle = \frac{\sigma^2}{N} + \langle \bar{X} \rangle^2.$$

Ma allora $\langle v^2 \rangle = \sigma^2 + m^2 - \sigma^2/N - m^2 = \sigma^2 - \sigma^2/N$. La miglior stima della varianza è data semplicemente da

$$\sigma^2 = \frac{1}{N-1} \sum_i (X_i - \bar{X})^2,$$

e questo $N - 1$ anziché N a denominatore è dovuto al fatto che abbiamo usato un grado di libertà per stimare la media.



2 Distribuzioni di probabilità univariate

Cominciamo questa sezione con qualche esempio di distribuzioni comuni, che verranno approfondite ulteriormente nella sezione. Un primo esempio comune è quello del dado, con $P_k = 1/6$ e $k = 1, \dots, 6$. In questo caso

$$\langle X \rangle = \sum_{k=1}^6 X_k P_k = \sum_{k=1}^6 k P_k = \frac{1}{6} \frac{(6+1) \times 6}{2} = \frac{7}{2}.$$

Per il secondo momento, analogamente,

$$\langle X^2 \rangle = \sum_{k=1}^6 k^2 P_k = \frac{1}{6} (1 + 4 + 9 + 16 + 25 + 36) = \frac{91}{6}.$$

E quindi $\sigma^2 = 91/6 - 49/4 = 35/12 \simeq (1.7)^2$, le “barre d'errore” del dado sono poco più piccole di 2, per cui $X = 3.5 \pm 1.7$. La skewness del dado è $\gamma_1 = 0$, per cui la PDF è simmetrica.

Un esempio più interessante è la *lifetime distribution* dei decadimenti. Tipicamente è scritta come

$$\rho(t) = \frac{1}{\tau} e^{-t/\tau}, \quad t \geq 0.$$

In questo caso

$$\langle t \rangle = \int_0^\infty dt \frac{t}{\tau} e^{-t/\tau} = \tau.$$

Analogamente si dimostra che

$$\langle t^n \rangle = \int_0^\infty dt \frac{t^n}{\tau} e^{-t/\tau} = \text{per parti} = n! \tau^n.$$

Quindi $\langle t^2 \rangle = 2\tau^2$ e $\sigma^2 = 2\tau^2 - \tau^2 = \tau^2$. La skewness chiaramente non è nulla essendo la distribuzione asimmetrica, infatti

$$\gamma_1 = \frac{m_3 + 3m\sigma^2 - m^3}{\sigma^3} = \frac{6\tau^3 - 3\tau^3 - \tau^3}{\tau^3} = 2 \neq 0.$$

Un terzo esempio è la distribuzione uniforme $\rho(x) = 1$ per $0 < x < 1$, e $\rho(x) = 0$ altrimenti. In questo caso

$$\begin{aligned} \langle x \rangle &= \int_0^1 dx \rho(x) x = \frac{1}{2} \\ \langle x^2 \rangle &= \int_0^1 dx \rho(x) x^2 = \frac{1}{3}, \end{aligned}$$

e quindi $\sigma^2 = \langle x^2 \rangle - \langle x \rangle^2 = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}$.

Infine c'è la gaussiana, con

$$\rho(x) = \frac{1}{\sqrt{2\pi}s} \exp\left(-\frac{(x-x_0)^2}{2s^2}\right),$$

in cui i primi due momenti sono semplicemente $m = x_0$ e $\sigma^2 = s^2$. La caratteristica della gaussiana è che è definita solamente dai primi due momenti: tutti gli altri sono nulli: $\gamma_1 = 0$, $\beta_2 = 0$, cioè $\mu_{k \geq 2} = 0$.

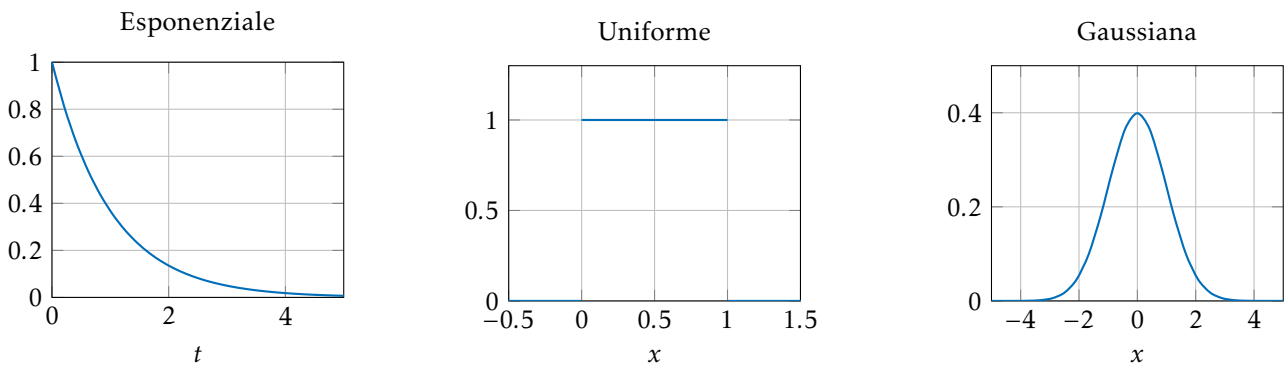


Figura 5: PDF di tre diverse distribuzioni (normalizzate): lifetime distribution con $\tau = 1$, uniforme, e gaussiana di media nulla $x_0 = 0$ e deviazione standard unitaria $s = 1$.



2.1 La funzione caratteristica

Il modo per generare questi momenti “in automatico”, senza integrali, è quello di sfruttare la funzione caratteristica (detta anche funzione generatrice), definita come la trasformata di Fourier della PDF:

$$\phi(t) := \langle e^{itx} \rangle = \int_{-\infty}^{\infty} dx e^{itx} \rho(x).$$

In caso di funzione discreta sarà ovviamente definita da una serie discreta di Fourier anziché un integrale:

$$\phi(t) := \sum_k e^{itx_k} P_k.$$

La funzione caratteristica ha diverse proprietà: innanzitutto, essendo la ρ normalizzata si ha che $|\phi(t)|^2 \leq 1$. Inoltre è appunto normalizzata, quindi $\phi(0) = 1$. Infine, una proprietà tipica della trasformata di Fourier è che $\phi(-t) = \phi^*(t)$ (coniugato), e quindi sia ha che per distribuzioni simmetriche vale $\rho(x) = \rho(-x) \implies \phi \in \mathbb{R}$.

Per quanto riguarda la trasformata inversa:

$$\rho(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} dt e^{-itx} \phi(t),$$

e nel caso discreto

$$P_k = \lim_{T \rightarrow \infty} \int_{-T}^T dt \phi(t) e^{-itx_k}.$$

Torniamo a parlare dei momenti. Questi li possiamo ottenere a partire dalla funzione caratteristica attraverso

$$\left. \frac{d^n \phi}{dt^n} \right|_{t=0} = \int_{-\infty}^{\infty} dx (ix)^n e^{itx} \rho(x) = i^n \int_{-\infty}^{\infty} dx x^n \rho(x) = i^n \mu_n \implies \boxed{\mu_n = (-i)^n \left. \frac{d^n \phi(t)}{dt^n} \right|_{t=0}}.$$

Quindi i primi due momenti (media e varianza) sono definiti da

$$\mu_1 = -i \left. \frac{d\phi(t)}{dt} \right|_{t=0} \quad \mu_2 = - \left. \frac{d^2 \phi(t)}{dt^2} \right|_{t=0}.$$

Possiamo espandere in serie di Taylor la funzione caratteristica attorno a 0 per ottenerne un'espressione analitica alternativa in funzione dei momenti stessi, ottenendo

$$\phi(t) = \sum_{n=0}^{\infty} \frac{t^n}{n!} \left. \frac{d^n \phi}{dt^n} \right|_{t=0} = \sum_{n=0}^{\infty} \frac{1}{n!} (it)^n \mu_n.$$

Per quanto riguarda i momenti centrali, li avevamo definiti come $\mu'_n = \langle (x - \mu)^n \rangle$. Quale sarà la funzione generatrice di questi momenti centrali? Sarà data da

$$\phi'(t) = \langle e^{it(x-\mu)} \rangle = e^{-it\mu} \phi(t).$$

Possiamo vedere che, analogamente a prima,

$$\left. \frac{d^n \phi'}{dt^n} \right|_{t=0} = i^n \mu'_n.$$

Di conseguenza, i primi due momenti centrali sono definiti da

$$\mu'_1 = -i \left. \frac{d\phi'(t)}{dt} \right|_{t=0} \quad \mu'_2 = - \left. \frac{d^2 \phi'(t)}{dt^2} \right|_{t=0}.$$

Questo è utile inoltre nel caso della convoluzione: consideriamo due variabili random indipendenti x e y . Definiamo la nuova distribuzione che definisce $z = x + y$. Cerchiamo la $\rho(z)$, integrando su tutti i valori di x, y con il vincolo sulla somma $z = x + y$

$$\rho(z) = \iint dx dy g(x) h(y) \delta(z - x - y) = \int dx g(x) h(z - x),$$



si ottiene una convoluzione delle singole PDF. La funzione caratteristica associata sarà data $\phi_\rho(t) = \phi_g(t)\phi_h(t)$. Consideriamo sempre il caso di una variabile random $z = x + y$ con distribuzioni $g(x)$, $h(y)$ e $f(z)$. Altra variazione sul tema dei momenti è l'argomento dei cumulanti. I cumulanti sono un'alternativa ai momenti, nel senso che due PDF che condividono gli stessi momenti avranno di conseguenza gli stessi cumulanti, e vice versa. Sono un modo diverso di espandere la funzione caratteristica in serie. In particolare, essi sono definiti in termini del logaritmo della funzione generatrice

$$K(t) = \ln \phi(t) = \ln \langle e^{itx} \rangle = K_1 it + K_2 \frac{(it)^2}{2!} + K_3 \frac{(it)^3}{3!} + \dots$$

Visto che $\phi(0) = 1$, non c'è termine costante, e abbiamo che $K_1 = \mu_1$, $K_2 = \mu_2' = \sigma^2$, $K_3 = \mu_3'$ e $K_4 = \mu_4' - 3\mu_2'^2$: i primi tre cumulanti sono uguali ai rispettivi momenti, dal quarto in poi inizia ad esserci una distinzione. I vari K_i sono detti cumulanti della distribuzione. A partire da questi si può ottenere la skewness della distribuzione come $\gamma_1 = K_3/K_2^{3/2}$.

Un'osservazione: in fisica statistica ci si riporta spesso al calcolo della **funzione di partizione** Z , che **non è altro che la funzione generatrice delle varie energie** (potenziali termodinamici). Si mostra che l'energia interna E è legata ai cumulanti: $K_1 = -\partial \ln Z / \partial \beta = \langle E \rangle$. Analogamente il calore specifico si lega al secondo cumulante $K_2 = \partial^2 \ln Z / \partial \beta^2 = k_B T c_V$.

Un primo esempio di applicazione è la distribuzione di Poisson

$$\mathcal{P}_\lambda(k) = \frac{\lambda^k}{k!} e^{-\lambda}.$$

La funzione caratteristica è

$$\phi(t) = \sum_k e^{itk} \frac{\lambda^k}{k!} e^{-\lambda} = e^{-\lambda} \sum_k \frac{(\lambda e^{it})^k}{k!} = \exp(\lambda e^{it} - \lambda).$$

A questo punto possiamo calcolare i momenti in maniera semplice: quello che occorre fare è semplicemente la derivata della funzione caratteristica e porre $t = 0$.

$$-i \left. \frac{d\phi}{dt} \right|_{t=0} = -i \left[\lambda i e^{it} e^{\lambda(e^{it}-1)} \right]_{t=0} = \lambda \implies \mu = \lambda.$$

Per il secondo momento otteniamo invece

$$- \left. \frac{d^2 \phi}{dt^2} \right|_{t=0} = - \left[e^{\lambda(e^{it}-1)} \left((\lambda i e^{it})^2 - \lambda e^{it} \right) \right]_{t=0} = \lambda^2 + \lambda = \mu_2 \implies \sigma^2 = \lambda = \mu.$$

Per Poisson abbiamo quindi che la varianza è uguale al valor medio! Altro fatto interessante sono i cumulanti:

$$K(t) = \lambda(e^{it} - 1) = \lambda \sum_{n=1}^{\infty} \frac{(it)^n}{n!} = \sum_{n=0}^{\infty} \frac{(it)^n}{n!} K_n,$$

e quindi tutti i cumulanti sono uguali: $K_1 = K_2 = \dots = K_n = \lambda$.

2.2 Trasformazione di variabili

Parliamo ora di trasformazione di variabili: il punto di partenza è una variabile casuale x distribuita con una certa $\rho(x)$. Immaginiamo di avere una certa funzione $u(x)$: qual è la PDF di questa funzione? Ci chiediamo quale sia la probabilità di avere questa funzione tra due valori \rightarrow sfruttiamo la conservazione della probabilità $P(x_1 < x < x_2) = 1 \implies P(u_1 < u < u_2)$ dove $u_1 = u(x_1)$ e $u_2 = u(x_2)$. Possiamo dire che

$$P(x_1 < x < x_2) = \int_{x_1}^{x_2} dx' \rho(x') = \int_{u_1}^{u_2} du g(u).$$

Cambiando variabili otteniamo lo jacobiano della distribuzione:

$$g(u) = \left| \frac{dx}{du} \right| \rho(x(u)).$$



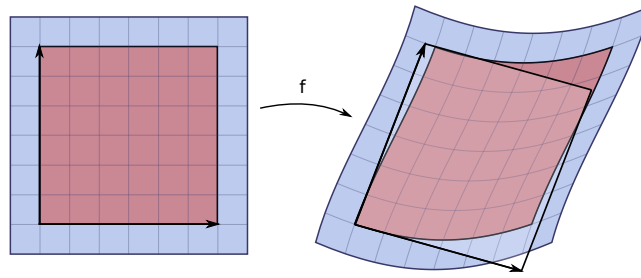


Figura 6: Esempio di una mappa non lineare $f : \mathbb{R}^2 \mapsto \mathbb{R}^2$ manda un piccolo quadrato in un parallelogramma distorto. Lo jacobiano dà la migliore approssimazione lineare del parallelogramma distorto, e il suo determinante dà il rapporto tra l'area del parallelogramma e quella del quadrato originario.

Facciamo un esempio: consideriamo delle sfere random, con raggio distribuito tra r_1 ed r_2 , e PDF uniforme $\rho(r) = 1/(r_2 - r_1)$ in questo intervallo (e zero altrimenti, cfr. Fig. 5 in centro). Qual è la distribuzione dei volumi di queste sfere $g(V)$? Sappiamo che $V = 4\pi r^3/3$. Usando la trasformazione otteniamo $g(V) = |dr/dV|\rho(r(V))$, con $r(V) = 3/(4\pi V^{1/3})$. Calcoliamo ora $|dr/dV| = 1/(4\pi V^{2/3})$, e quindi

$$g(V) = \frac{1}{V_2^{1/3} - V_1^{1/3}} \frac{1}{3} V^{-2/3}.$$

Vogliamo adesso determinare la trasformazione da una variabile casuale ad un'altra. L'idea è che abbiamo due variabili casuali e vogliamo sapere la trasformazione che mandi $x \mapsto y$. Utilizziamo il fatto che le distribuzioni cumulate devono essere uguali—si devono conservare i “volumi” di probabilità:

$$F(x) = \underbrace{\int_{-\infty}^x dx' f(x')}_{\text{conservazione volumi}} = \int_{-\infty}^y dy' g(y') = G(y).$$

Quindi $F(x) = G(y) \implies y(x) = G^{-1}(F(x))$, che vale solo se la G è una funzione invertibile. Prendendo l'esempio del primo notebook di Python, vogliamo passare da una distribuzione di probabilità $f(x)$ uniforme ad una $g(y)$ esponenziale:

$$f(0 \leq x \leq 1) = 1 \quad \text{e} \quad g(y \geq 0) = \lambda e^{-\lambda y}.$$

Quindi, integrando da $-\infty$ all'estremo superiore di integrazione, otteniamo la trasformazione

$$\left. \begin{aligned} G(y) &= \int_0^y dy' \lambda e^{-\lambda y'} = -e^{-\lambda y'} \Big|_0^y = 1 - e^{-\lambda y} \\ F(x) &= \int_0^x dx' = x \end{aligned} \right\} \implies x = 1 - e^{-\lambda y} \implies y = -\frac{1}{\lambda} \ln(1 - x).$$

2.3 Calcolo combinatorio

Il calcolo combinatorio non fa propriamente parte della probabilità, tuttavia tutti i calcoli di probabilità discrete hanno a che fare in qualche modo con il calcolo combinatorio. Adoperando la definizione classica di probabilità infatti dobbiamo contare i casi favorevoli su quelli totali.

Un primo esempio che possiamo fare è considerare delle coppie di elementi statisticamente indipendenti. Tipicamente, se abbiamo due coppie in cui il primo elemento può avere n elementi e il secondo m , allora il numero di coppie possibili è $N_p = n \times m$. Il passo successivo è considerare i multiplotti, semplice generalizzazione di questo: il numero totale di multiplotti $N_m = n_1 \times n_2 \times \dots \times n_r = \prod_{i=1}^r n_i$. Questo si dimostra per induzione a partire dalla dimostrazione fatta “contando” per le coppie. Un esempio di applicazione di calcolo combinatorio semplice è quello del modello di Ising, in cui abbiamo N spin che possono prendere valore $s_i = \pm 1$. La domanda più semplice che ci possiamo fare è quante siano le configurazioni possibili \rightarrow multiplotto. In questo caso $n_1 = n_2 = \dots = n_N = 2$ e quindi applicando la regola del multiplotto, il numero totale di configurazioni del modello di Ising è 2^N .

Se abbiamo un dado con cui facciamo r tiri, qual è la probabilità che esca *almeno un 6*? Per ogni tiro, la probabilità che esca un 6 è pari a $1/6$, e se facciamo r tiri, la probabilità che non esca è $5/6$. Quindi la probabilità che non esca nessun 6 sarà $(5/6)^r$, e quindi la soluzione è $P_r = 1 - (5/6)^r$.



Supponiamo adesso di avere n campioni ordinati a_1, \dots, a_n , e di fare delle estrazioni a_{j_1}, \dots, a_{j_r} . Quante sono le possibili estrazioni, i modi in cui posso prendere questi elementi? Ci sono due modi di fare queste estrazioni: la prima è quella di farle “senza rimpiazzo” (estriamo dal mazzo senza rimetterla dentro). In questo modo ad ogni scelta successiva il ventaglio delle possibilità diminuisce. L'altro caso è quella delle estrazioni “con rimpiazzo”, in cui l'elemento viene rimesso nel mazzo.

Un esempio dell'effetto delle estrazioni senza rimpiazzo è quello del Blackjack, gioco di carte in cui il mazzo si assottiglia estrazione per estrazione, riuscendo a far contare ai giocatori più abili le carte uscite e poter fare puntate più sicure. Infatti, se in un mazzo di 52 carte mano dopo mano queste non vengono rimpiazzate, è sempre più probabile che una carta non ancora uscita venga estratta dal mazzo.

Tornando al mazzo, quante sono le possibilità distinte considerando n di fare r estrazioni? Nel caso con rimpiazzo il mazzo non cambia e ho sempre le stesse possibilità, per cui torniamo ad n^r . Nel primo caso (estrazioni senza rimpiazzo) invece abbiamo n possibilità per la prima estrazione, $n-1$ per la seconda e così via \rightarrow facendo il prodotto di tutti fattoriali otteniamo $n!/(n-r)!$. Osserviamo che per $r = n$ estraiamo tutte le carte e quindi abbiamo $n!$ possibilità (permutazioni).

Esercizio: tirando un dado sei volte, qual è la probabilità di avere tutti e sei gli esiti diversi?

Basta contare il numero di possibilità ad ogni tiro: al primo va bene qualunque numero, al secondo ci sono cinque possibilità buone su sei, al terzo quattro e così via. Le probabilità non sono indipendenti, per cui vanno moltiplicate secondo

$$P = 1 \times \frac{5}{6} \times \frac{4}{6} \times \frac{3}{6} \times \frac{2}{6} \times \frac{1}{6} = \frac{120}{7776} \approx 1.5\%.$$

Rivediamo il risultato dal punto di vista classico. In questo caso avremmo delle estrazioni con rimpiazzo, $P = n/n_{\text{tot}} = 6!/6^6$ che porta allo stesso risultato, senza aver contato!

Altro punto importante sono le partizioni. Una partizione è una divisione di una popolazione di oggetti in parti. Due partizioni sono diverse se almeno un elemento è diverso. Ci chiediamo quindi qual è la probabilità di una sottopopolazione di r elementi tra n . Quante possibilità distinte di avere cinque carte in un mazzo da sessantaquattro? In questo caso non conta l'ordine, per cui il numero di possibilità indipendenti è

$$N_p = \frac{n!}{(n-r)!r!} = \binom{n}{r},$$

e dividiamo per $r!$ perché non vogliamo contare più volte lo stesso gruppo di carte. Il termine tra parentesi è detto coefficiente binomiale e ha alcune proprietà utili: innanzitutto $\binom{n}{0} = 1$. Per $r > n$ si ha che $\binom{n}{r} = 0$. Infine vale che $\binom{n}{r} = \binom{n}{n-r}$.

Un'applicazione utile è il teorema binomiale: visto che dobbiamo fare il prodotto di un certo numero di oggetti (coppie), si ha che

$$(x+y)^n = \sum_{k=0}^n \binom{n}{k} x^k y^{n-k}.$$

Continuiamo con il calcolo combinatorio. Siamo in 33 in questa aula virtuale. Ci chiediamo quale sia la probabilità che ci sia qualcuno tra di noi nato la data odierna? Sarà alta o bassa? È una coincidenza rara o probabile che due persone abbiano il compleanno lo stesso giorno?

Esercizio: qual è la probabilità che, dato un insieme di N persone, almeno una di queste sia nata oggi? E quella che almeno due siano nate lo stesso giorno?

Conviene maneggiare le probabilità complementari per questo tipo di esercizi. La probabilità di non essere nato oggi è, assumendo una PDF uniforme, $\bar{P}_N = \frac{364}{365}$ per una persona. Per N persone avremo $\left(\frac{364}{365}\right)^N$, cioè la probabilità che nessuno sia nato oggi. Per avere la probabilità che *almeno uno* sia nato oggi la probabilità sarà quindi

$$P_N = 1 - \bar{P}_N = 1 - \left(\frac{364}{365}\right)^N = 1 - \left(1 - \frac{1}{365}\right)^N \approx 1 - \left(1 - \frac{N}{365}\right) = \frac{N}{365}.$$

Anche il secondo punto conviene ragionare analogamente. Possiamo vedere la probabilità che nessuno



sia nato lo stesso giorno come un'estrazione senza rimpiazzo da un mazzo di carte:

$$\bar{P}_N = \underbrace{\left(1 - \frac{1}{365}\right)}_{1 \text{ confronto}} \underbrace{\left(1 - \frac{2}{365}\right)}_{2 \text{ confronti}} \underbrace{\left(1 - \frac{3}{365}\right)}_{3 \text{ confronti}} \cdots \left(1 - \frac{N-1}{365}\right).$$

Possiamo dare una forma più elegante a $\bar{P}_N = \frac{365 \times 364 \times \cdots \times (365 - N + 1)}{(365)^N} = \frac{365!}{(365 - N)! 365^N}$. Sfruttando il coefficiente binomiale è

$$\bar{P}_N = \frac{N!}{365^N} \binom{365}{N}.$$

Possiamo ulteriormente semplificare l'espressione per \bar{P}_N sfruttando l'espansione al primo ordine dell'esponenziale $e^x \approx 1 - x$ per $x \ll 1$:

$$\bar{P}_N \approx \prod_{k=1}^{N-1} e^{-\frac{k}{365}} = \exp\left(-\sum_{k=1}^{N-1} \frac{k}{365}\right) = \exp\left(-\frac{N(N-1)}{730}\right) \Rightarrow P_N \approx 1 - e^{-\frac{N(N-1)}{730}} \approx 1 - e^{-N^2/730}.$$

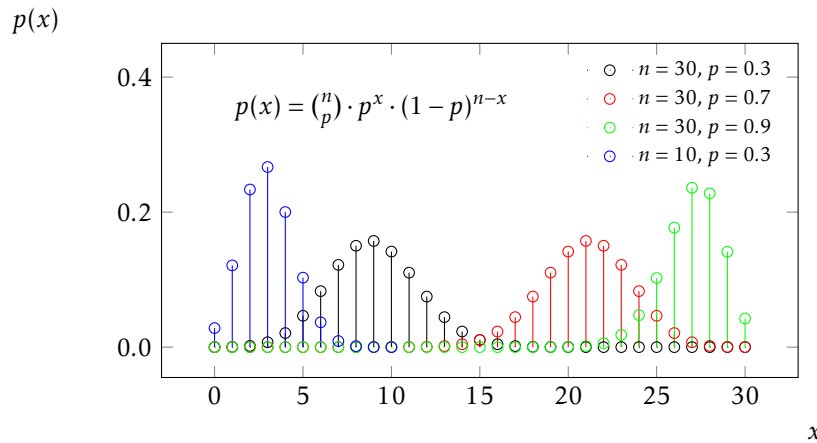


Figura 7: PDF della distribuzione binomiale. Si osservi come si allarghi la distribuzione all'aumentare del numero di campioni n (a parità di probabilità).

2.3.1 Distribuzione binomiale

Alcuni esperimenti consistono nell'eseguire ripetutamente una data operazione, come lanciare più volte una moneta o un dado. In qualche modo, la probabilità non varierà da una prova all'altra per via dell'indipendenza di queste.

La distribuzione binomiale è quella distribuzione che descrive problemi in cui abbiamo più estrazioni con rimpiazzo di un certo evento. Se le estrazioni sono fatte senza rimpiazzo, queste non sono indipendenti e si ottiene la distribuzione ipergeometrica. In ogni caso, se $N \gg n$ numero di estrazioni, la binomiale resta una buona approssimazione. Supponiamo di tirare un certo numero di dadi, facciamo dieci. Qual è la probabilità di avere due volte 6? La probabilità che esca 6 è $p = 1/6$, mentre $q = 1 - p = 5/6$. Perché escano due 6 su 10 tiri la probabilità è data da $p^2(1-p)^8$ ma questo non basta, perché dobbiamo considerare tutti i possibili scambi con cui possiamo estrarre i due risultati vincenti su un totale di 10. Questo viene dato dal coefficiente binomiale, per cui abbiamo $\binom{10}{2} p^2 (1-p)^8$. Questa è la distribuzione binomiale. Generalizzando il risultato di prima per k successi in n eventi, se p è la probabilità di successo e $q = 1 - p$ la probabilità che l'evento desiderato non avvenga, la probabilità che l'evento avvenga k volte è data da

$$\mathcal{B}_p^n(k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

Come ci mostra la Fig. 7, la binomiale è una distribuzione discreta (la distribuzione "continua" è un limite per $n \rightarrow \infty$), ed è normalizzata, cioè

$$\sum_{k=0}^n \mathcal{B}_p^n(k) = 1$$



per via del teorema del binomio di Newton

$$(x + y)^n = \sum_{k=0}^n \binom{n}{k} x^k y^{n-k} = (p + 1 - p)^n = 1,$$

dove i due termini elevati alla n sono le rispettive probabilità di successo e insuccesso. Ovviamente la somma di questi due esempi esclusivi restituisce lo spazio intero della probabilità. Quali sono i momenti? In teoria dovremmo applicare la formula

$$\langle k \rangle = \sum_{k=0}^n k \binom{n}{k} p^k (1-p)^{n-k},$$

ma il modo più efficiente di ottenerli è quello di sfruttare la funzione caratteristica, visto che calcolare esplicitamente tale sommatoria è un compito complesso. Questa era definita per una distribuzione discreta come la serie di Fourier della distribuzione stessa:

$$\begin{aligned} \phi(t) &= \sum_{k=0}^n e^{ikt} \mathcal{B}_p^n(k) = \sum_{k=0}^n e^{ikt} \binom{n}{k} p^k (1-p)^{n-k} = \sum_{k=0}^n \binom{n}{k} (e^{it} p)^k (1-p)^{n-k} \\ &= (e^{it} p + 1 - p)^n, \end{aligned}$$

dove per passare dalla prima alla seconda riga abbiamo sfruttato il teorema del binomio di Newton con $x = e^{it} p$ e $y = (1 - p)$. Di conseguenza il primo momento sarà

$$\langle k \rangle = \mu = -i \left. \frac{d\phi}{dt} \right|_{t=0} = np e^{it} [1 + p(e^{it} - 1)]^{n-1} \Big|_{t=0} = np.$$

Derivando un'altra volta rispetto a t questa espressione si trova il secondo momento:

$$\langle k^2 \rangle = - \left. \frac{d^2 \phi}{dt^2} \right|_{t=0} = e^{it} np [1 + p(e^{it} - 1)]^{n-1} \Big|_{t=0} + e^{it} np^2 e^{it} [1 + p(e^{it} - 1)]^{n-2} \Big|_{t=0} = np + n(n-1)p^2.$$

Per la varianza si ha di conseguenza che per la binomiale vale

$$\langle k^2 \rangle - \langle k \rangle^2 = n(n-1)p^2 + np - n^2 p^2 = np(1-p) = \sigma^2.$$

C'è anche un'alternativa per calcolare questi momenti, sfruttando il teorema di Newton per cui

$$(p + q)^n = \sum_{k=0}^n \binom{n}{k} p^k q^{n-k} =: f(p, q).$$

Osserviamo che $f(p, 1-p) = 1$ per via della binomiale. Inoltre

$$p \left. \frac{\partial f}{\partial p} \right|_{q=1-p} = \sum_k k p^k q^{n-k} \binom{n}{k} \equiv \langle k \rangle.$$

Stessa cosa poi per il secondo momento in cui

$$\langle k^2 \rangle = p \frac{\partial}{\partial p} p \left. \frac{\partial f}{\partial p} \right|_{q=1-p}.$$

2.3.2 Distribuzione multinomiale

La generalizzazione della distribuzione binomiale è la distribuzione multinomiale, utile nel caso ci siano più risultati possibili, in generale con probabilità diverse \rightarrow non avremo binomi ma multinomi, e rimpiazzeremo il coefficiente binomiale con quello multinomiale, dato dal numero di partizioni in k classi (la distribuzione binomiale era $k = 2$ e $n > 1$, per $k = 2$ e $n = 1$ si ottiene la distribuzione di Bernoulli) di dimensione n_1, n_2, \dots, n_k tali che $\sum_{i=1}^k n_i = N$. Questo numero $N(n_1, \dots, n_k) = N! / n_1! n_2! \dots n_k!$. Anche qui vale un teorema simile a quello di Newton ma per il multinomio:

$$(a_1 + a_2 + \dots + a_k)^N = \sum_{n_1, \dots, n_k} \frac{N!}{\prod_i n_i!} a_1^{n_1} \dots a_k^{n_k}.$$



Dalla PDF della distribuzione binomiale possiamo facilmente a quella multinomiale, che è

$$p(n_1, \dots, n_k) = \frac{N!}{n_1! \dots n_k!} p_1^{n_1} \dots p_k^{n_k} = N! \prod_{i=1}^k \frac{p_i^{n_i}}{n_i!}.$$

La possiamo vedere come una specie di binomiale in cui ci sono più di due probabilità. Un esempio di applicazione della multinomiale è studiare un campione di 10 persone e vederne il gruppo sanguigno. Supponendo che il gruppo 0 si presenti statisticamente nella popolazione con il 10% di probabilità, quelli A e B con il 35% l'uno e quello AB con il restante 20%, con quale probabilità avremo 3 persone con gruppo 0, 2 con A, 4 con B e 1 con AB? Per rispondere a questa domanda si sfrutta appunto il coefficiente multinomiale. La distribuzione multinomiale ha diverse proprietà. Il primo momento è simile a quello della binomiale, e lo si ottiene sfruttando il teorema multinomiale:

$$\left(\sum_i p_i \right)^N = 1 = \sum_{n_1, \dots, n_k} p(n_1, \dots, n_k) = f(p_1, \dots, p_k).$$

Facendo la derivata e sommandola su p_i otteniamo quindi il primo momento i -esimo

$$\langle n_i \rangle = p_i \frac{\partial}{\partial p_i} f = N p_i \left(\sum_i p_i \right)^{N-1} \Big|_{\sum p_i=1} = N p_i = \langle n_i \rangle.$$

La covarianza si ottiene a partire da

$$\begin{aligned} \langle n_i n_j \rangle &= p_i \frac{\partial}{\partial p_i} p_j \frac{\partial}{\partial p_j} f = p_i \frac{\partial}{\partial p_i} p_j \frac{\partial}{\partial p_j} \left(\sum_j p_j \right)^N \\ &= p_i \frac{\partial}{\partial p_i} N p_j \left(\sum_j p_j \right)^{N-1} \quad \text{derivata prodotto, } \partial p_j / \partial p_i = \delta_{ij} \\ &= p_i N \delta_{ij} \left(\sum_j p_j \right)^{N-1} + N p_i p_j (N-1) \left(\sum_j p_j \right)^{N-2} \\ &= N p_i \delta_{ij} + N(N-1) p_i p_j \quad \text{nella multinomiale la covarianza ha } i \neq j \\ &= N(N-1) p_i p_j \end{aligned}$$

e quindi

$$\text{Cov}(n_i, n_j) = \langle n_i n_j \rangle - \langle n_i \rangle \langle n_j \rangle = N(N-1) p_i p_j - \cancel{N p_i} \times \cancel{N p_j} = -N p_i p_j.$$

Vogliamo calcolare la varianza della multinomiale imponendo $p_i = p_j$. Calcoliamo come prima cosa il secondo momento per indici uguali:

$$\begin{aligned} \langle n_i^2 \rangle &= p_i \frac{\partial}{\partial p_i} p_i \frac{\partial}{\partial p_i} f = p_i \frac{\partial}{\partial p_i} N p_i \left(\sum_i p_i \right)^{N-1} = p_i N \left(\sum_i p_i \right)^{N-1} + p_i N(N-1) p_i \left(\sum_i p_i \right)^{N-2} \\ &= p_i N + p_i^2 N(N-1) \\ &= p_i^2 N^2 + N p_i (1 - p_i) = \langle n_i^2 \rangle. \end{aligned}$$

Di conseguenza, ricordando che $\langle n_i \rangle = N p_i$, la varianza si ottiene come

$$\text{Var} = \text{Cov}_{i=j} = \langle n_i^2 \rangle - \langle n_i \rangle^2 = N p_i (1 - p_i).$$

Di conseguenza $\text{Cov}_{i \neq j} = -N p_i p_j$.

2.4 Distribuzione di Poisson

La distribuzione di Poisson è una distribuzione di probabilità che esprime la probabilità che un dato numero di eventi occorranza ad intervalli temporali fissati. Questo processo avviene con un rate λ —probabilità per



unità di tempo—indipendentemente dalla scala di tempo considerata. Qual è la probabilità che avvengano in questa unità di tempo n eventi? In generale questa probabilità è data dalla distribuzione di Poisson

$$\mathcal{P}_n(\lambda) = \frac{e^{-\lambda} \lambda^n}{n!}.$$

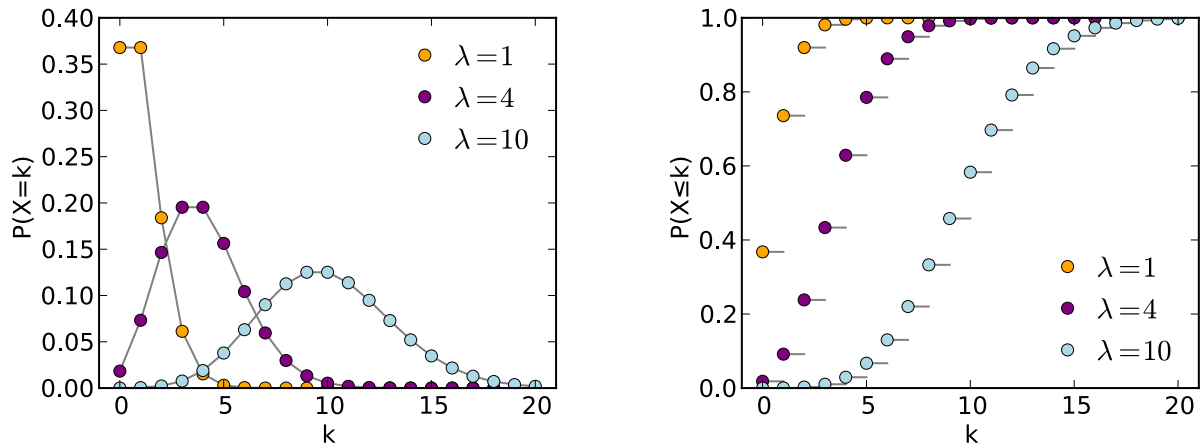


Figura 8: PDF e cumulativa di più distribuzioni poissoniane a confronto.

Possiamo anche vederla come limite continuo della distribuzione binomiale (cfr. Fig. 9):

$$\lim_{\substack{N \rightarrow \infty \\ Np = \text{const} = \lambda}} \mathcal{B}_p^N(n) = \mathcal{P}_n(\lambda).$$

Adesso il numero di tiri non è su un tempo discreto ma su un intervallo di tempo continuo. Questo si può vedere in vari modi. Il primo è, come già detto, fare il limite della binomiale sfruttando Stirling, i limiti $N \gg n$, $N \rightarrow \infty$ e la condizione $Np = \lambda$:

$$\begin{aligned} \binom{N}{n} p^n (1-p)^{N-n} &\approx \frac{N^N e^{-N} N^n}{(N-n)^{N-n} e^{-N+n} n!} p^n (1-p)^{N-n} = \frac{N^N \left(\frac{\lambda}{N}\right)^n \left(1 - \frac{\lambda}{N}\right)^{N-n} e^{-n}}{N^{N-n} \left(1 - \frac{n}{N}\right)^{N-n} n!} \\ &= \frac{\lambda^n \left(1 - \frac{\lambda}{N}\right)^{N-n} e^{-n}}{\left(1 - \frac{n}{N}\right)^{N-n} n!} \underset{N \gg n}{\approx} \frac{\lambda^n \left(1 - \frac{\lambda}{N}\right)^N e^{-n}}{\left(1 - \frac{n}{N}\right)^N n!}. \end{aligned}$$

Ricordandosi il limite notevole $\lim_{n \rightarrow \infty} (1 - x/n)^n = e^{-x}$,

$$\binom{N}{n} p^n (1-p)^{N-n} \approx \frac{\lambda^n e^{-\lambda} e^{-n}}{e^{-n} n!} = \frac{\lambda^n e^{-\lambda}}{n!}.$$

C'è una dimostrazione più semplice che sfrutta la funzione caratteristica della binomiale. Per la binomiale avevamo $\phi_B(t) = (1 + p(e^{it} - 1))^N$. Facendo lo stesso limite di prima $Np = \lambda$,

$$\lim_{\substack{N \rightarrow \infty \\ Np = \lambda}} \phi_B(t) = \lim_{N \rightarrow \infty} \left(1 + \frac{\lambda(e^{it} - 1)}{N}\right)^N = \exp(\lambda(e^{it} - 1)) = \phi_P(t).$$

Si può in alternativa calcolare esplicitamente la funzione caratteristica:

$$\begin{aligned} \phi_P(t) &= \sum_n e^{itn} p_n = \sum_n e^{itn} \frac{\lambda^n}{n!} e^{-\lambda} = e^{-\lambda} \sum_n \frac{\lambda^n}{n!} e^{itn} = e^{-\lambda} \exp(\lambda e^{it}) \\ &= \exp(\lambda(e^{it} - 1)). \end{aligned}$$



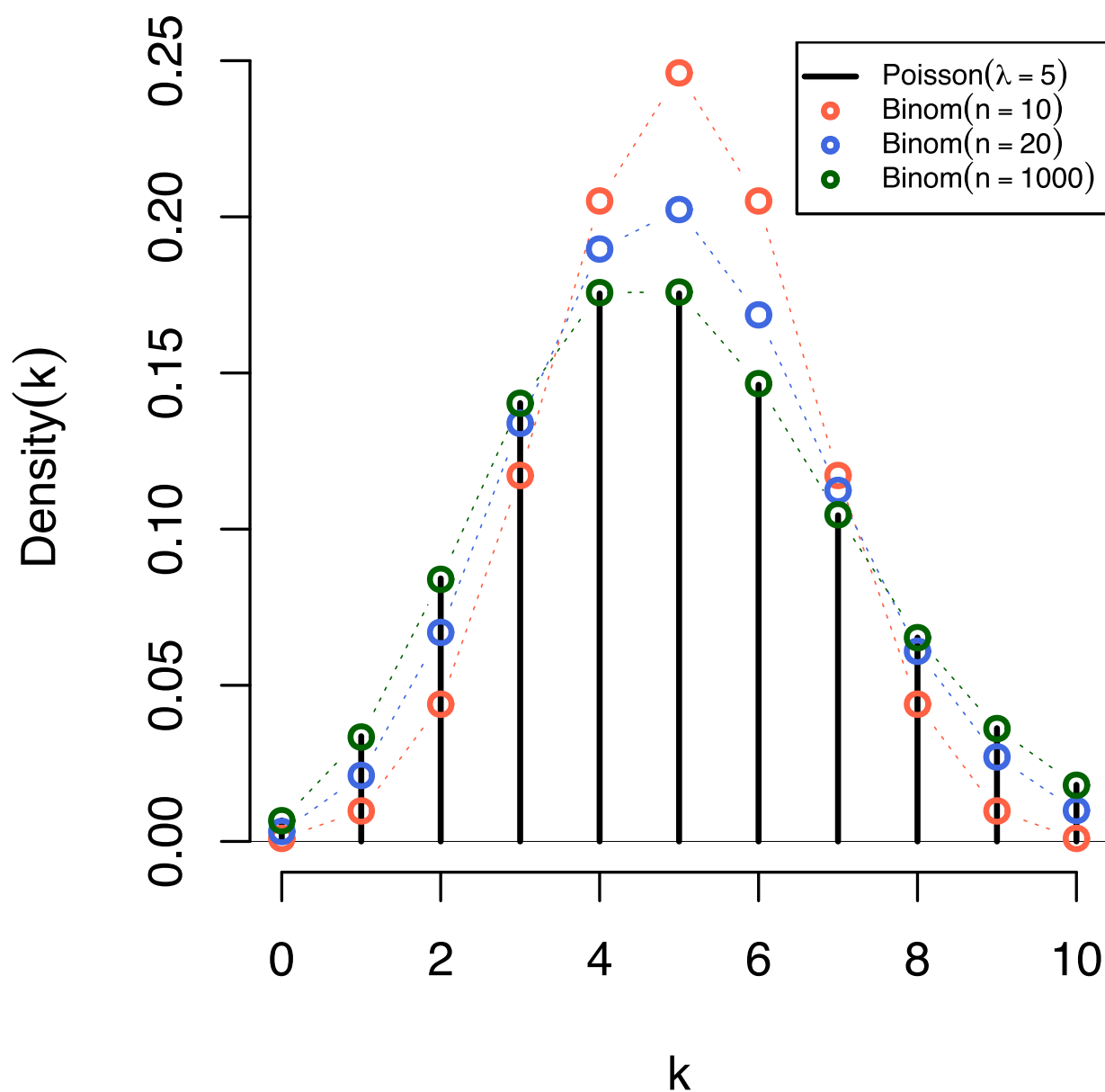


Figura 9: Distribuzione di Poisson come limite della distribuzione binomiale al variare del numero di tiri.

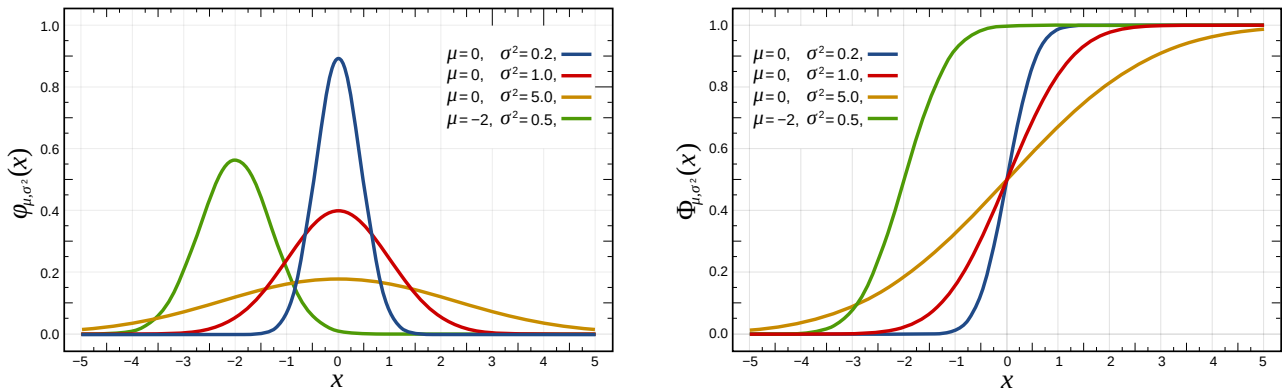


Figura 10: PDF e cumulativa di più distribuzioni gaussiane a confronto.

A partire dalla funzione caratteristica possiamo calcolare i momenti della distribuzione di Poisson:

$$\left. \begin{aligned} \langle n \rangle &= -i \left. \frac{d\phi_P(t)}{dt} \right|_{t=0} = -i \lambda i e^{it} e^{\lambda(e^{it}-1)} \Big|_{t=0} = \lambda \\ \langle n^2 \rangle &= - \left. \frac{d^2\phi_P(t)}{dt^2} \right|_{t=0} = \left[\lambda e^{it} e^{\lambda(e^{it}-1)} + \lambda^2 (e^{it})^2 e^{\lambda(e^{it}-1)} \right]_{t=0} = \lambda + \lambda^2 \end{aligned} \right\} \Rightarrow \sigma^2 = \lambda.$$

Cosa succede quando sommiamo due distribuzioni di Poisson? Immaginiamo due processi con parametri rispettivamente λ_1 e λ_2 . Supponendoli indipendenti avremo n_1 ed n_2 eventi. Sommiamo i due processi poissoniani indipendenti $\Rightarrow n = n_1 + n_2$. Sommeremo su tutti i possibili eventi compatibili sul constraint su n :

$$\mathcal{P}_n = \sum_{n_1, n_2} \delta_{n_1, n_1+n_2} \mathcal{P}_{n_1} \mathcal{P}_{n_2}.$$

Essendo una somma su convoluzioni conviene trasformare con Fourier, per cui la convoluzione diventa un prodotto:

$$\phi(t, \lambda) = \phi_P(t, \lambda_1) \phi_P(t, \lambda_2) = \phi_P(t, \lambda_1 + \lambda_2).$$

Questo è un segno di “stabilità”: la somma di due poissoniane è una poissoniana con rate sommati—un po’ come succede per le gaussiane, anch’esse distribuzioni “stabili”.

2.5 Distribuzione Gaussiana

La distribuzione gaussiana è

$$\mathcal{N}(x|\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}},$$

dove $\langle x \rangle = \mu$ e $\langle x^2 \rangle - \langle x \rangle^2 = \sigma^2$. Sempre sfruttando la convoluzione si può dimostrare che $x = x_1 + x_2$ è gaussiana, di media $\mu = \mu_1 + \mu_2$ e varianza $\sigma^2 = \sigma_1^2 + \sigma_2^2$. La funzione caratteristica di una gaussiana è un integrale gaussiano (e quindi anch’essa una gaussiana) che si risolve completando il quadrato sotto il segno di integrale

$$\phi(t) = \int_{-\infty}^{\infty} dx e^{-\frac{(x-\mu)^2}{2\sigma^2} + ixt} = \exp\left(-\sigma^2 \frac{t^2}{2} + i\mu t\right).$$

La ϕ della somma è uguale a

$$\phi(t) = \exp\left(-\sigma_1^2 \frac{t^2}{2} + i\mu_1 t - \sigma_2^2 \frac{t^2}{2} + i\mu_2 t\right) = \exp\left(-\frac{\sigma_1^2 + \sigma_2^2}{2} t^2 + i(\mu_1 + \mu_2)t\right).$$

Il logaritmo di questa ci permette di trovare i cumulanti con opportune derivazioni:

$$\ln \phi = \frac{\sigma^2}{2} t^2 + i\mu t.$$



Si osserva che i primi due cumulanti sono gli unici a non essere nulli, infatti tutte le derivate di ordine pari o superiore a tre sono nulle, essendo $\ln \phi$ una funzione parabolica in t :

$$K_1 = \mu \quad K_2 = \sigma^2 \quad K_{\geq 3} = 0.$$

Di conseguenza i primi due momenti saranno non nulli coincidendo con i cumulanti, così come invece saranno pari a zero tutti i momenti di ordine superiore al secondo.

2.6 Distribuzione del χ^2

Supponiamo di avere x_i, σ_i dove $i = 1, \dots, k$ e $\langle x_i \rangle = 0$. Quando facciamo una misurazione, il valore che abbiamo misurato sarà quello vero a meno di fluttuazioni dovute alla misura stessa. La distribuzione del χ^2 , definita da $\chi^2 = \sum_{i=1}^k x_i^2 / \sigma_i^2 \Rightarrow \sum_{i=1}^k (x_i - \mu_i)^2 / \sigma_i^2$ ci permette di capire quanto buona è la nostra misura.

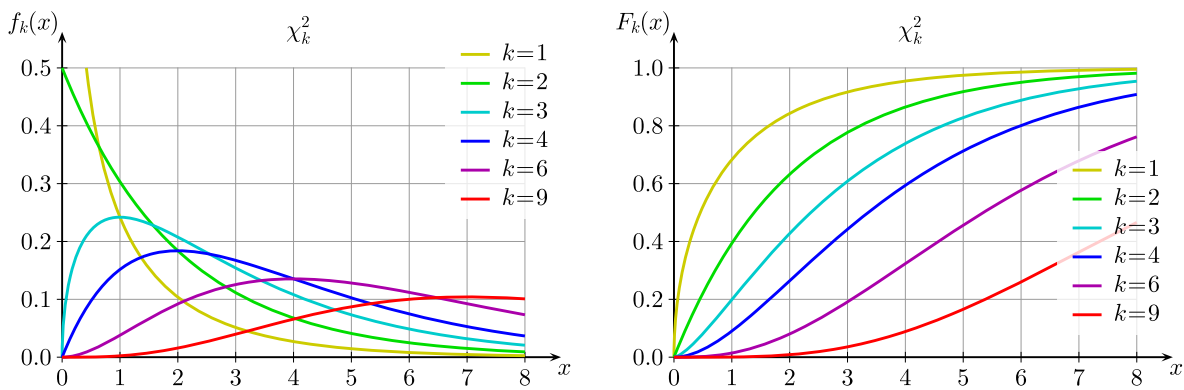


Figura 11: Distribuzione del χ^2 e rispettive cumulative.

Facciamo il caso più semplice $k = 1$. Sia $u = x^2 / \sigma^2$ il nostro chi quadro monodimensionale. Supponiamo che la nostra x abbia distribuzione normale $\rho(x) = N(x|0, \sigma)$. Scriviamo quindi la PDF della distribuzione:

$$g_1(u) = \rho(u) \left| \frac{dx}{du} \right|,$$

dove $x = \pm \sigma \sqrt{u}$ e quindi

$$\left| \frac{dx}{du} \right| = \frac{\sigma}{2\sqrt{u}} \Rightarrow g_1(u) = \frac{1}{\sqrt{2\pi u}} e^{-u/2}.$$

Nel caso generale $k \neq 1$ questa è più complessa, e prende la forma:

$$g_k(u) = \frac{u^{\frac{k}{2}-1} e^{-u/2}}{\Gamma(\frac{k}{2}) 2^{\frac{k}{2}}},$$

dove

$$\Gamma(z) = \int_0^\infty dt t^{z-1} e^{-t},$$

è la funzione Gamma di Eulero, tale che $\Gamma(n+1) = n!$. Nell'analisi statistica degli errori ricorre spesso la funzione del chi quadro:

$$\chi^2 = \sum_{i=1}^N \frac{(x_i - x_i^{\text{teo}}(\lambda_1, \dots, \lambda_Z))^2}{\sigma_i^2},$$

e si dovrà minimizzare rispetto ai vari λ .

2.7 Distribuzione Gamma

La distribuzione Gamma è una distribuzione di probabilità continua, che comprende, come casi particolari, anche le distribuzioni esponenziale e chi quadrato. Viene utilizzata come modello generale dei tempi di attesa nella teoria delle code, soprattutto qualora siano importanti effetti che rimuovano "l'assenza di memoria"



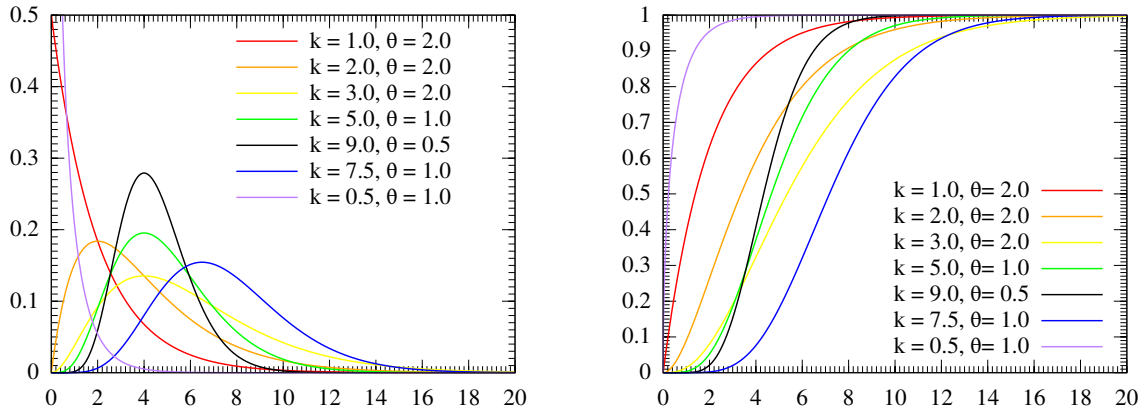


Figura 12: Alcune distribuzioni Gamma e rispettive cumulative, dove $\nu = k$ e $\lambda = 1/\theta$.

della distribuzione esponenziale. Nella statistica bayesiana è comune sia come distribuzione a priori che come distribuzione a posteriori. Essa è definita dalla PDF

$$G(x|\nu, \lambda) = \frac{\lambda^\nu}{\Gamma(\nu)} x^{\nu-1} e^{-\lambda x}, \quad x > 0.$$

In questo caso λ = fattore di scala, e ν = forma della distribuzione. Osserviamo che per $\lambda = 1/2$ e $\nu = k/2$ otteniamo la distribuzione del chi quadro, mentre per $\nu = 1$ otteniamo la distribuzione esponenziale $G = \lambda e^{-\lambda x}$. È simile alla distribuzione di Poisson. È importante per questo suo essere una “super-distribuzione” che racchiude altre distribuzioni caratteristiche come casi particolari.

La funzione caratteristica è difficile da calcolare, ma si dimostra essere pari a

$$\phi(t) = \left(1 - \frac{it}{\lambda}\right)^{-\nu},$$

da cui si possono calcolare i momenti.

$$\left. \begin{aligned} \langle x \rangle &= -i \frac{d\phi}{dt} \Big|_{t=0} = -i^2 \frac{\nu}{\lambda} \left(1 - \frac{it}{\lambda}\right)^{-\nu+1} = \frac{\nu}{\lambda} \\ \langle x^2 \rangle &= -\frac{d^2\phi}{dt^2} \Big|_{t=0} = \frac{d}{dt} \left[\frac{iv}{\lambda} \left(1 - \frac{it}{\lambda}\right)^{-\nu-1} \right]_{t=0} = -\left[\frac{iv}{\lambda} \frac{i(\nu+1)}{\lambda} \left(1 - \frac{it}{\lambda}\right) \right]_{t=0} = \frac{\nu^2}{\lambda^2} + \frac{\nu}{\lambda^2} \end{aligned} \right\} \Rightarrow \sigma^2 = \frac{\nu}{\lambda^2},$$

simile alla distribuzione di Poisson, con varianza simile a media. Il momento i -esimo si calcola facilmente, ed è uguale a

$$\mu_i = \frac{\Gamma(i + \nu)}{\lambda^i \Gamma(\nu)}.$$

Immaginiamo di avere due variabili x_1 e x_2 , distribuite rispettivamente con parametri $G(x_1, \nu_1)$ e $G(x_2, \nu_2)$. Cerchiamo $\rho(x = x_1 + x_2)$. La distribuzione della somma si può scrivere come

$$\rho(x = x_1 + x_2) = \int_{-\infty}^{\infty} dz G(z, \nu_1) G(z - x, \nu_2)$$

e quindi

$$\phi(t) = \phi_G(t, \nu_1) \phi_G(t, \nu_2) = \left(1 - \frac{it}{\lambda}\right)^{-\nu_1} \left(1 - \frac{it}{\lambda}\right)^{-\nu_2} = G(x|\nu_1 + \nu_2).$$

Questa distribuzione può essere sfruttata per descrivere il decadimento di particelle, che hanno distribuzione esponenziale.

2.8 Distribuzione Beta

La distribuzione beta è una distribuzione che vale nell'intervallo $[0, 1]$. Abbiamo una dipendenza da due parametri:

$$p_\beta(x|\alpha, \rho) = \frac{x^{\alpha-1} (1-x)^{\rho-1}}{B(\alpha, \rho)}, \quad B(\alpha, \rho) = \int_0^1 dt t^{\alpha-1} (1-t)^{\rho-1} = \frac{\Gamma(\alpha)\Gamma(\rho)}{\Gamma(\alpha+\rho)}.$$



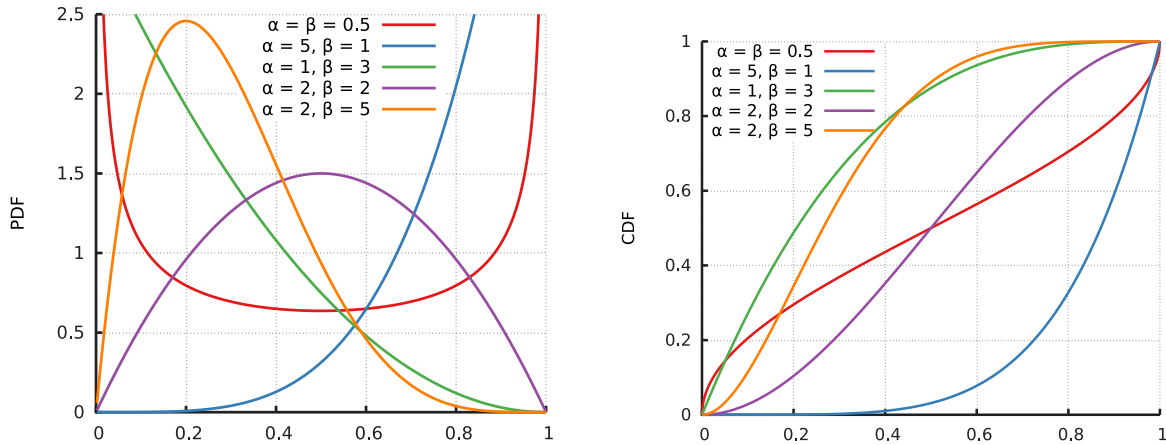


Figura 13: Distribuzioni Beta e rispettive cumulative.

Questo parametro è la generalizzazione del coefficiente binomiale, cui si riconduce per determinati valori:

$$\frac{1}{B(n+1, N-n+1)} = \frac{\Gamma(n+2)}{\Gamma(n+1)\Gamma(N-n+1)} = (n+1) \frac{N!}{n!(N-n)!} = (n+1) \binom{N}{n}.$$

Osserviamo lo stretto rapporto con la binomiale $\mathcal{B}_p^N(k) = p_\beta(p|k+1, N-k+1)$. Le due distribuzioni tuttavia descrivono processi leggermente diversi, nella distribuzione binomiale p è un parametro (la probabilità dell'evento) nella distribuzione Beta p è invece una variabile casuale.

È interessante per studiare la statistica di ordinamento (*order statistics*): immaginiamo una serie di variabili random distribuite i.i.d., cioè identiche e distribuite indipendentemente le une dalle altre. Vediamo la cumulata $\rho(x) \mapsto F(x)$, ordinando le variabili dalla più piccola alla più grande s_1, \dots, s_L tali che $s_1 < \dots < s_L$, sono ordinate. Ci chiediamo come sia distribuita la k -esima variabile. Qual è la probabilità $p(s_k \in [x, x+dx])$? Questa densità di probabilità è quella di estrarre prima $k-1$ numeri minori di x , ed $L-k$ maggiori di x . Sfruttando le cumulative otteniamo:

$$p(s_k \in [x, x+dx]) = \frac{F(x)^{k-1}(1-F(x))^{L-k}}{B(k, L-k+1)} \rho(x) dx$$

Ovviamente bisogna fare attenzione agli scambi, da cui il fattore a denominatore che non è altro che un fattore binomiale. Qual è $F(x)$? Se prendiamo come $\rho(x)$ una distribuzione uniforme $\rho(x) = 1$ tra $[0, 1]$ allora $F(x) = x$ e quindi

$$p(s_k = k) = \frac{x^{k-1}(1-x)^{L-k}}{B(k, L-k+1)} \equiv p_\beta(x|k, L-k),$$

e osserviamo quindi come la distribuzione Beta capiti naturalmente nell'ordinamento di variabili random i.i.d..

2.9 Distribuzione di Cauchy

La distribuzione di Cauchy

$$p_C(x) = \frac{1}{\pi(1+(x-x_0)^2)},$$

è simmetrica rispetto alla media x_0 . Nonostante abbia una forma a campana come la Gaussiana, non ha varianza finita per via delle code "lunghe". La cosa interessante è che i primi due momenti divergono in \mathbb{R} . La varianza in particolare è infinita, per cui la distribuzione ha senso solo in presenza di un cutoff, dal valore massimo che può avere x . La funzione caratteristica per $x_0 = 0$ diventa $\phi(t) = \exp(-|t|)$, che ovviamente non è derivabile in $t = 0$.

2.10 Distribuzione di Lévy

Un'altra distribuzione che ha momenti divergenti è quella di Lévy, che non ha una forma analitica precisa. Infatti è descritta dalla sua funzione caratteristica,

$$\phi_L(z) = \exp\left(-c_0 + ic_i \frac{z}{|z|} |z|^\alpha\right).$$



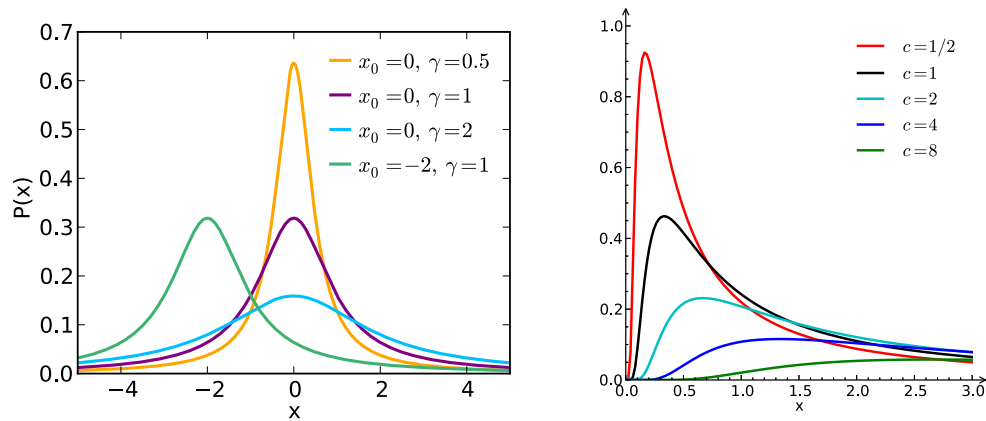


Figura 14: Distribuzioni di Cauchy e Levy a confronto.

Caso particolare $c_i = 0$ genera

$$\phi_L(z) = e^{-c_0|z|^\alpha},$$

che coincide con una gaussiana per $\alpha = 2$. Solitamente però $\alpha < 2$, e quindi la funzione caratteristica ha code divergenti e varianza divergente:

$$p_L(x) = \frac{1}{2\pi} \int dz \phi_L(z) e^{ixz}.$$

Le code sono “larghe”, decadono a potenza. Si osserva quindi per processi invarianti per scala. Inoltre è una distribuzione stabile: la somma di variabili di tipo Levy sono variabili di tipo Levy, così come avviene per gaussiane e gamma. Inoltre ha una proprietà in più: il bacino di attrazione.

Riprendiamo il discorso del calcolo combinatorio di qualche lezione fa con qualche estensione. Ci poniamo questo problema, quello del poker Texas hold 'em. Noi abbiamo due carte che conosciamo, e sul tavolo ci sono cinque carte—tre scoperte e due coperte. Con queste sette carte qui possiamo sceglierne cinque. Le carte si scoprono dopo ogni puntata. La domanda che ci poniamo è la seguente: vogliamo cercare di fare colore (cinque carte dello stesso seme). Supponiamo di avere in mano ($4\heartsuit, J\heartsuit$), che sul tavolo ci siano invece ($2\heartsuit, 5\heartsuit, A\spadesuit$). Qual è la probabilità che tra queste due carte coperte ci sia il seme mancante? Se la probabilità è alta conviene puntare, altrimenti non puntiamo ovviamente. Ci sono disponibili ancora 47 carte, e ci servono le probabilità di estrarre una o due carte di cuori tra queste carte qui. Esiste in realtà una distribuzione di probabilità che descrive questa situazione, ed è la distribuzione ipergeometrica.

Facciamo prima un passo indietro, e pensiamo alle probabilità—o meglio, ai numeri—di occupazione, tipiche della fisica quantistica statistica (con distribuzioni di bosoni e fermioni).

2.11 Numeri di occupazione

Abbiamo una serie di oggetti $\vec{n} = (n_1, \dots, n_k)$ in k scatole. Sappiamo che il numero totale di oggetti è fissato $\sum_i n_i = N$. Qual è il modo in cui possiamo distribuire queste particelle in queste scatole?

2.11.1 Bosoni

In che modo possiamo scambiare particelle e partizioni considerando tutte le possibilità? Il numero di modi in cui possiamo distribuire N oggetti in k scatole è dato da

$$A_{N,k} = \binom{N+k-1}{N} \Rightarrow P_{N,k} = \frac{1}{A_{N,k}}.$$

Quindi la probabilità di ogni configurazione è semplicemente $\frac{1}{A_{N,k}}$. Questa è la statistica di Bose. Fondamentale è il fatto che gli oggetti siano *indistinguibili*. Se invece sono distinguibili ritroviamo la statistica classica, in cui ogni particella ha una probabilità di occupare la scatola j data da $p_j = 1/k$ perché sono tutti uguali e quindi abbiamo una multinomiale $p_n = n! / (\sum_j n_j k^n)$.



2.11.2 Fermioni

Per i fermioni servono tante scatole: $k \geq N$ —lasciamo per ora perdere lo spin. Ci servono N scatole occupate e $k - N$ vuote. In questo caso la probabilità è data da

$$P_N = \frac{1}{\binom{k}{N}}.$$

Torniamo a parlare delle distribuzioni.

2.12 Distribuzioni geometrica e ipergeometrica

Sono distribuzioni che si usano per estrazioni di palline da un sacco, carte da un mazzo e simili, e inoltre hanno applicazioni anche in ambiti più “esotici” come l’ecologia. Supponiamo di avere n oggetti, di cui n_1 rossi e $n_2 = n - n_1$ neri. Ci poniamo il problema di estrarre in maniera sequenziale gli oggetti, uno dopo l’altro. Con k estrazioni, ci chiediamo quale sia la probabilità che esca “nero” per la prima volta alla k -esima estrazione. Dobbiamo distinguere due possibilità: *con* e *senza* rimpiazzo.

Analizziamo prima il caso *senza* rimpiazzo: ad ogni istante (estrazione) ci saranno meno palline. Per rispondere alla domanda, per tutte le altre $k - 1$ estrazioni deve essere uscito “rosso”. In questo caso abbiamo quindi che la probabilità di estrarre rosse è

$$\frac{n_1!}{(n - (k - 1))!},$$

numero di modi in cui posso estrarre palline rosse su $k - 1$ estrazioni. Quando siamo alla k -esima estrazione vogliamo tirare fuori una pallina nera, e idealmente ne ho “tante”. Quanti modi ho di estrarle? Ne ho n_2 e devo normalizzare rispetto alle palline che ho e al numero totale di eventi possibili $n!/(n - k)!$:

$$p = \frac{\frac{n_1!}{(n_1 - (k - 1))!} n_2}{\frac{n!}{(n - k)!}},$$

che possiamo riscrivere come

$$\frac{n_1! n_2 (n - k)!}{(n_1 - k + 1)! n!}.$$

Se invece andiamo a vedere il caso *con* rimpiazzo, il numero totale di eventi è $n_{\text{tot}} = n^k$. La probabilità che esca una particella rossa è $p_1 = n_1/n$, mentre $p_2 = n_2/n$. La probabilità con rimpiazzo è quindi data semplicemente da

$$p = p_1^{k-1} p_2 = \left(\frac{n_1}{n}\right)^{k-1} \frac{n_2}{n}.$$

Quest’ultima è nota come distribuzione geometrica. Rimarchiamo come questa sia connessa con le estrazioni con rimpiazzo. Viene utilizzata per descrivere la probabilità di successo di un processo di Bernoulli dopo un numero k di insuccessi. Ad esempio può essere usata per descrivere la probabilità che esca testa dopo 1, 2, o 13 lanci di moneta. Il valor medio della distribuzione geometrica è

$$\langle k \rangle = \sum_k p_1 (1 - p_1)$$

dove, considerando il limite infinito, si sfrutta la serie geometrica

$$\sum_{k=0}^{\infty} p_1^k = \frac{1}{1 - p_1}$$

da cui il nome. La PDF è data da

$$p(k) = p(1 - p)^k.$$

Possiamo calcolare questo p_1 sfruttando la solita tecnica come

$$p_1 \frac{\partial}{\partial p_1} \left(\sum_{k=0}^{\infty} p_1^k \right) = \sum_k k p_1^k = p_1 \frac{\partial}{\partial p_1} \left(\frac{1}{1 - p_1} \right) = \frac{p_1}{(1 - p_1)^2} \implies \langle k_1 \rangle = \frac{p_1}{1 - p_1}.$$



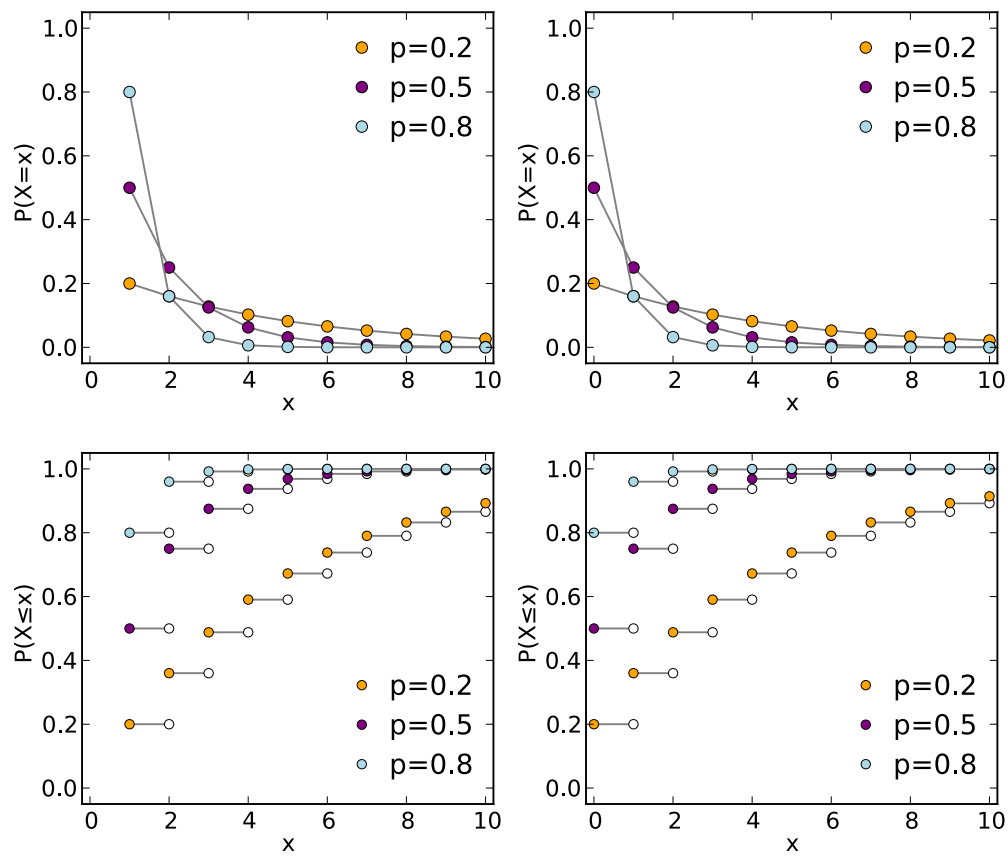


Figura 15: Distribuzioni geometriche e rispettive cumulative.

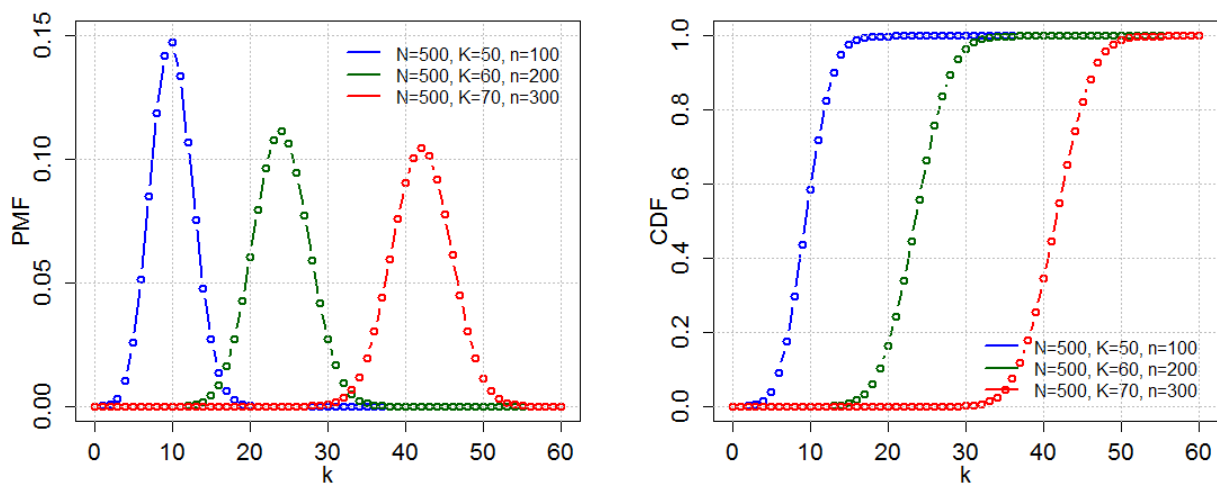


Figura 16: Distribuzioni ipergeometriche e relative cumulative.



Analogamente si dimostra che

$$\text{Var}(k) = \frac{p_1}{(1-p_1)^2}.$$

Veniamo ora a parlare della distribuzione ipergeometrica. Il problema è leggermente diverso: estraiamo sempre k oggetti su un totale di $n = n_1 + n_2$. Vogliamo sapere la probabilità che k_1 siano rossi e k_2 siano neri, con il vincolo $k = k_1 + k_2$. Per tornare al discorso delle carte, supponiamo estrazioni senza rimpiazzo. Il numero totale di possibili estrazioni è $\binom{n}{k}$, indipendentemente da come estraggo le particelle. Poi abbiamo il numero di estrazioni “di tipo 1” (cioè le estrazioni di palline rosse): quante possibilità ho di farlo sulle palline rosse? Sarà $\binom{n_1}{k_1}$, e similmente per estrazioni “di tipo 2” avremo $\binom{n_2}{k_2}$. Di conseguenza

$$P(k_1|k, n_1, n_2) = \frac{\binom{n_1}{k_1}\binom{n_2}{k_2}}{\binom{n}{k}} = \frac{\binom{n_1}{k_1}\binom{n_2}{k_2}}{\binom{n_1+n_2}{k_1+k_2}}.$$

Questa è la PDF della distribuzione ipergeometrica (Fig. 16).

Torniamo al discorso della partita di poker di prima. Qual è la probabilità di fare colore vista la disposizione di carte precedente? Abbiamo considerato il caso in cui si estraggano due carte che non siano cuori, e le sottraiamo all'unità—dal momento che ci interessa che esca *almeno una* carta di cuori. Basta sostituire nell'ipergeometrica $n_1 = 9$ numero di carte di cuori rimaste, $n_2 = 38$ numero carte non di cuori rimanenti, $k_1 = 0$ e $k_2 = 2$ —non estraiamo nessuna delle 9 carte di cuori, ed estraiamone 2 tra le 38 non di cuori—, con quindi una probabilità circa del 35%.

Illustriamo adesso un apparentemente singolare collegamento tra il gioco del poker e lo screening di una popolazione animale. Come fanno gli ecologi a stimare le popolazioni degli animali? Chiaramente è impensabile poterli contare tutti quando si tratta di (decine/centinaia di) migliaia di individui, quindi bisogna usare dei metodi statistici o probabilistici. Uno dei metodi più standard è basato appunto sulla distribuzione ipergeometrica. Supponiamo che vogliamo stimare il numero di pesci in un lago. L'idea è fare delle estrazioni e sfruttare la distribuzione ipergeometrica. Supponiamo ci siano n pesci, e ne peschiamo al tempo $t = 0$ un numero n_r marcandoli, mettendoci un label e ributtandoli nel bacino. Aspettiamo un po' di tempo, e facciamo una nuova pesca prendendo k pesci al $t = 1$. In questo istante di tempo conteremo quanti pesci saranno marcati k_r . A questo punto possiamo marcare la probabilità che ci siano n pesci nel lago (ragionamento Bayesiano) $P(n|k_r, k, n_r)$. Sfruttando il teorema di Bayes,

$$P(n|k_r, k, n_r) = \frac{1}{Z} P(k_r|n, n_r, k) P(n|k, n_r).$$

Possiamo usare la funzione ipergeometrica per calcolare

$$P(k_r|n, n_r, k) = \frac{\binom{n_r}{k_r}\binom{n-n_r}{k-k_r}}{\binom{n}{k}}.$$

Ci siamo quasi, ci serve capire cosa è $P(n|k, n_r)$ probabilità di avere n pesci indipendentemente da tutto il resto (sarebbe il prior Bayesiano). Facciamo quindi delle ipotesi di “massima ignoranza”: sicuramente $n \geq n_r$, e inoltre $n < n_{\max} = V_{\text{lago}}/V_{\text{pesce}}$ per esempio. Possiamo quindi scrivere $P(n|k, n_r) = 1/n_{\max} - n_r$ per $n \in [n_r, n_{\max}]$ e zero altrimenti (step function). Inserendo tutto con la normalizzazione, troviamo che dato k_r e k , possiamo andare a vedere come è fatta la distribuzione. Una volta marcati x pesci dopo la prima pesca, se pesco tante volte un numero di pesci marcati vicino a x è molto probabile che ci siano poco più di x pesci nel lago. Se quando ri-pesco invece ne ho pochi di pesci marcati, la stima del numero totale di pesci è ovviamente molto più ampia.

$$P(k_1|k, n_1, n_2) = \frac{\binom{n_1}{k_1}\binom{n_2}{k_2}}{\binom{n}{k}}$$

Il fatto che la distribuzione ipergeometrica sia intimamente connessa con le estrazioni senza rimpiazzo e il campionamento di una popolazione come quella dei pesci appena fatta, è dovuto al fatto che quando peschiamo dei pesci “nuovi”, non marcati, li segniamo e li togliamo dalla popolazione dei pesci ignoti.

Ci sono tante applicazioni anche in ambito genomico. Vediamo che in un malato c'è un numero maggiore di determinati geni di altri. I geni sono solitamente organizzati in cluster ontologici, in gruppi cooperanti per una certa funzione \rightarrow pathways che indicano i gruppi di geni cooperanti. Supponiamo di aver trovato un certo numero di geni fuori dal normale: sono associati a qualche funzione descritta da questi pathways? Questi geni sono lì per caso o perché c'è veramente una funzione che è stata sregolata? Qui entra in gioco la distribuzione ipergeometrica: qual è la probabilità di prendere k geni sregolati e metterli in quel determinato cluster?



Il valor medio e la varianza dell'ipergeometrica sono

$$\langle k_1 \rangle = n_1 \frac{k}{n}$$

$$\sigma^2 = \frac{k(n-k)n_1n_2}{(n-1)n^2},$$

ma sono piuttosto complicati da dimostrare.

Ultimo caso da menzionare è quello in cui si ha estrazione di k oggetti (di due tipi k_1 e k_2) da un campione di n elementi. Consideriamo il caso con rimpiazzo questa volta. In questo caso la probabilità è, considerando tutti gli scambi,

$$p(k_1) = \binom{n_1}{n}^{k_1} \binom{n_2}{n}^{k-k_1} \binom{k}{k_1} \Rightarrow \text{binomiale} \quad \mathcal{B}_{p_1}^k(k_1) = p_1^{k_1} p_2^{k-k_1} \binom{k}{k_1},$$

si torna al binomiale nel caso con rimpiazzo.

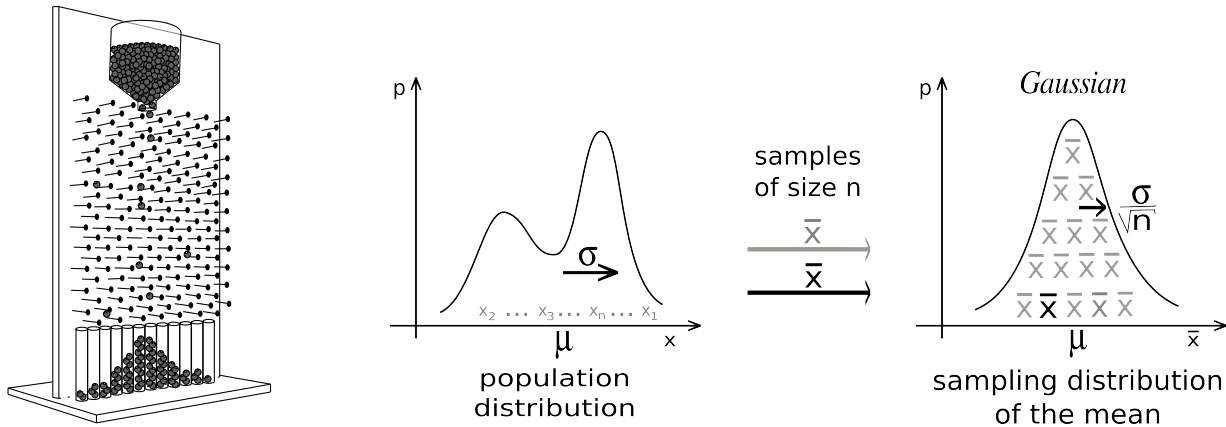


Figura 17: Rappresentazione del teorema del limite centrale. A sinistra, il classico esempio in cui cadono una serie di biglie con pari probabilità di andare a destra o sinistra ad ogni bivio: la distribuzione risultante quando vengono collezionate viene approssimata sempre più precisamente da una gaussiana mano a mano che le biglie cadono.

3 Il teorema del limite centrale

Questo teorema dà la risposta a quale sia la distribuzione di una variabile random data dalla somma di altre variabili random. Consideriamo una serie di variabili $x_i = 1, \dots, N$ i.i.d., in generale con una distribuzione $\rho(x_i)$ con varianza $\text{Var}(x_i) = \sigma_x^2 < \infty$ (il fatto che sia finita è fondamentale perché sia valido il CLT). Scriviamo la somma divisa per il numero di variabili (media)

$$s = \frac{1}{N} \sum_{i=1}^N x_i.$$

Il teorema del limite centrale (CLT) afferma che

$$\lim_{N \rightarrow \infty} P_N(s) = \mathcal{N}(s | \mu_s, \sigma_s),$$

valido per qualunque distribuzione di partenza di varianza finita. In questo caso

$$\mu_s = \mu_x = \langle x_i \rangle, \quad \sigma_s^2 = \sigma_x^2 / N,$$

dove l'uguaglianza della varianza è dovuta alla divisione per N nella definizione di y .

Il CLT può essere dimostrato a partire da una qualunque PDF $\rho(x)$ che caratterizza le variabili random i.i.d.. Consideriamo inizialmente il caso (più semplice) dove le variabili x_i sono estratte da una distribuzione gaussiana $\rho(x) = \mathcal{N}(x | \mu_x, \sigma_x)$. Sia $y_i = \frac{1}{N}(x_i - \mu_x)$. Allora

$$s_y = \sum_i y_i \quad \text{e} \quad s_x = \frac{1}{N} \sum_i x_i = s_y + \mu_x.$$



Per dimostrare questo teorema conviene considerare la funzione caratteristica, che nel caso delle variabili gaussiane (la somma di due variabili i.i.d. gaussiane è anch'essa gaussiana, distribuzione stabile!) y_i è anch'essa una gaussiana della forma

$$\phi_{y_i}(t) = \exp\left(-\frac{1}{2}\sigma_y^2 t^2\right),$$

perchè la media è nulla e quindi non c'è il termine complesso $i\mu_y t$. Infatti, come avevamo detto precedentemente, per distribuzioni simmetriche rispetto all'origine, la funzione caratteristica è reale—e nel caso della gaussiana è a sua volta una gaussiana. La funzione caratteristica della somma di due variabili è pari al prodotto delle singole funzioni caratteristiche

$$\phi_{y_1+y_2}(t) = \phi_{y_1}(t)\phi_{y_2}(t).$$

Per N variabili, di conseguenza, la funzione caratteristica è data dal prodotto delle N funzioni caratteristiche, e in generale sarà pari a

$$\phi_{s_y} = [\phi_y(t)]^N = \exp\left(-\frac{N}{2}\sigma_y^2 t^2\right) = \exp\left(-\frac{\sigma_x^2}{2N} t^2\right).$$

Si vede che la funzione caratteristica è quella di una gaussiana, con una nuova varianza riscalata

$$\sigma_{s_y}^2 = \frac{\sigma_x^2}{N}.$$

Quindi la distribuzione di s_y è una distribuzione normale con valor medio nullo e varianza riscalata

$$p(s_y) = \mathcal{N}(s_y|0, \sigma_x/\sqrt{N}).$$

Per arrivare al risultato del CLT basta riscrivere in funzione di s_x traslando la media:

$$p(s_x) = \mathcal{N}(s_x|\mu_x, \sigma_x/\sqrt{N}).$$

Il rapporto tra le varianze della variabile somma e quella delle singole variabili i.i.d. è valido $\forall N$, e non solo asintoticamente.

Consideriamo adesso una distribuzione esponenziale, e vediamo che il CLT torna anche per questa. Per dimostrarlo, abbiamo quindi $\rho(x) = \lambda e^{-\lambda x}$ con $x > 0$. Sia $s = \sum_i x_i$, e quindi vale $\phi_s(t) = (\phi_x(t))^N$. Scrivendo esplicitamente,

$$\phi_x(t) = \int_0^\infty dx \lambda e^{-\lambda x + itx} = \frac{\lambda}{\lambda - it}.$$

Di conseguenza

$$\phi_s(t) = [\phi_x(t)]^N = \left(\frac{1}{1 - it/\lambda}\right)^N.$$

A noi però interessa la $\phi_{s/N}(t)$, cioè il valore della funzione caratteristica “della media”, in un certo senso. Valendo $\phi_{s/N}(t) = \phi_s(t/N)$, prendendo il valor medio di \bar{x} si ottiene

$$\phi_{\bar{x}}(t) = \left(\frac{1}{1 - it/\lambda N}\right)^N = \exp(-N \ln(1 - it/\lambda N)) \approx \exp\left(-N\left(-\frac{it}{\lambda N} - \frac{1}{2}\left(-\frac{it}{\lambda N}\right)^2\right)\right) = \exp\left(\frac{it}{\lambda} - \frac{t^2}{2\lambda N}\right),$$

siamo tornati alla caratteristica della gaussiana, con $\mu_{\bar{x}} = 1/\lambda$ e $\sigma_{\bar{x}}^2 = 1/N\lambda^2$.

Il CLT si può dimostrare anche nel caso di distribuzione uniforme, $P(x_i) = 1$ per $0 < x_i < 1$, e $s = \sum_{i=1}^N x_i$. Quindi

$$\phi(t) = \int_0^1 dx e^{itx} = \frac{e^{it} - 1}{it}.$$

Ovviamente vale sempre $\phi_s(t) = (\phi(t))^N$. Ci serve calcolare

$$p(s|N) = \frac{1}{2\pi} \int_{-\infty}^{\infty} dt \left(\frac{e^{it} - 1}{it}\right)^N e^{-itx} = [\dots] = \text{polinomio in } s.$$

Nel caso generale $s_y = \sum_i \frac{1}{N}(x_i - \mu_x) \implies \phi_{s_y}(t) = (\phi_y(t))^N = \exp(N \ln \phi_y(t))$. Espandendo per piccoli t otteniamo

$$\phi_y(t) \approx 1 - \frac{\sigma_x^2}{2N^2} t^2 \implies N \ln \phi_y(t) \approx -\frac{\sigma_x^2}{2N} t^2,$$



e quindi

$$\phi_{s_y}(t) \approx e^{-\frac{\sigma_y^2 t^2}{2N}}.$$

Questo è meno evidente nel caso in cui la distribuzione non abbia varianza finita (distribuzione di Lévy, Sec. 2.10). Questa era definita solamente dalla sua funzione caratteristica

$$\phi_L(t) = \exp(-\alpha_\mu |t|^\mu).$$

La distribuzione non ha funzione analitica nota (solo come integrale), ma nel limite $x \rightarrow \infty$ allora

$$L_\mu(x) \sim \frac{\mu}{|x|^{1+\mu}}.$$

Osserviamo che per $\mu < 2$ si ha che

$$\sigma^2 \approx \int^L dx x^2 L_\mu(x) \sim L^{2-\mu}, \quad L \text{ cutoff.}$$

Per $\mu < 2$ la varianza diverge. Nonostante questo, vale lo stesso una forma del teorema del limite centrale—che però non restituirà una gaussiana. Se sommo N variabili di Lévy, ognuna con la propria $\rho(x) = L_\mu(x)$, qual è $p(s)$? Essendo le variabili indipendenti, la funzione caratteristica sarà $\phi_s(t) = (\phi(t))^N = \exp(-\alpha_\mu N |t|^\mu)$. Questa è sempre una funzione caratteristica di Lévy, ma con un nuovo parametro $\alpha' = \alpha N$. Esiste un risultato in più: se abbiamo una qualsiasi distribuzione $p(x)$ con varianza divergente $\langle x^2 \rangle \rightarrow \infty$, se nel limite asintotico abbiamo decadimento delle code a potenza del tipo $p(x) \sim 1/|x|^{1+\mu}$, allora la distribuzione della somma è sempre una distribuzione di Lévy.

4 Distribuzioni di probabilità multivariate

Come primo esempio prendiamo quello di due variabili: immaginiamo di avere x e y . Esiste una funzione cumulata

$$F(x, y) = p(\{x' < x\} \wedge \{y' < y\}).$$

Devono valere delle proprietà di normalizzazione: $F(-\infty, \infty) = 1$ e $F(-\infty, y) = F(x, -\infty) = 0$. La densità di probabilità sarà data dalla derivata mista della cumulata, in analogia con quanto succede con le distribuzioni univariate

$$\rho(x, y) = \frac{\partial^2 F}{\partial x \partial y}, \quad \iint dx dy \rho(x, y) = 1.$$

Dalla PDF possiamo definire le distribuzioni marginali, ristrette ad una sola variabile:

$$\rho_x(x) = \int_{-\infty}^{\infty} dy \rho(x, y), \quad \rho_y(y) = \int_{-\infty}^{\infty} dx \rho(x, y).$$

Non dobbiamo confondere la PDF di due variabili con la probabilità Bayesiana condizionata, che è effettivamente una

$$P_x(x|y) = \frac{\rho(x, y)}{\rho_y(y)}.$$

Similmente possiamo chiamare

$$P_y(y|x) = \frac{\rho(x, y)}{\rho_x(x)},$$

riformulazione facilmente estendibile a più variabili. Possiamo quindi riscrivere il teorema di Bayes:

$$P_x(x|y)\rho_y(y) = P_y(y|x)\rho_x(x).$$

Date queste distribuzioni possiamo definire i momenti:

$$\begin{aligned} \mu_x = \langle x \rangle &= \int dx x \rho_x(x) & \mu_y = \langle y \rangle &= \int dy y \rho_y(y) \\ \sigma_x^2 &= \langle (x - \mu_x)^2 \rangle & \sigma_y^2 &= \langle (y - \mu_y)^2 \rangle \\ \sigma_{xy}^2 &= \langle (x - \mu_x)(y - \mu_y) \rangle. \end{aligned}$$



Invece la misura della correlazione delle variabili normalizzata ad 1 è espressa da

$$\rho_{xy} = \sigma_{xy} / \sigma_x \sigma_y, \quad |\rho_{xy}| \leq 1.$$

Un caso semplice è quello di variabili indipendenti: in tal caso $\rho(x, y) = \rho(x)\rho(y)$ e $\rho_{xy} = 0$, dal momento che

$$\langle (x - \mu_x)(y - \mu_y) \rangle = \langle (x - \mu_x) \rangle \langle (y - \mu_y) \rangle = 0.$$

Se volessimo trasformare dei multipletti di variabili $(x, y) \mapsto (u(x, y), v(x, y))$? Avremo due PDF $f(x, y)$ e $g(u, v)$, e generalizzeremo la formula ottenuta precedentemente per la conservazione dei volumi di probabilità

$$g(u, v) = f(x, y) \left| \frac{\partial(x, y)}{\partial(u, v)} \right|,$$

con il determinante dello jacobiano.

Ad esempio, consideriamo una distribuzione gaussiana

$$f(x, y) = \frac{1}{2\pi} e^{-\frac{x^2+y^2}{2}}.$$

Vogliamo mapparla in coordinate polari $(x = r \cos \varphi, y = r \sin \varphi)$. Quindi scriviamo innanzitutto il determinante dello jacobiano

$$\left| \frac{\partial(x, y)}{\partial(r, \varphi)} \right| = r,$$

infatti

$$\left| \frac{\partial(x, y)}{\partial(r, \varphi)} \right| = \det \begin{pmatrix} \frac{\partial x}{\partial r} & \frac{\partial x}{\partial \varphi} \\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \varphi} \end{pmatrix} = \det \begin{pmatrix} \cos \varphi & -r \sin \varphi \\ \sin \varphi & r \cos \varphi \end{pmatrix} = r \cos^2 \varphi + r \sin^2 \varphi = r,$$

come volevasi dimostrare. Quindi

$$g(r, \varphi) = \frac{1}{2\pi} r e^{-r^2/2},$$

e di conseguenza la distribuzione si fattorizza in $g_r(r) = r e^{-r^2/2}$, distribuzione di Rayleigh, e $g_\varphi(\varphi) = 1/2\pi$, distribuzione uniforme.

Vediamo adesso il caso generale con x_1, \dots, x_n . La cumulata sarà $F(x_1, \dots, x_n) = P(\{x'_1 < x_1\}, \dots, \{x'_n < x_n\})$. Quindi

$$\rho(x_1, \dots, x_n) = \frac{\partial^n F}{\partial x_1 \dots \partial x_n}.$$

Interessante è la matrice di covarianza

$$C_{ij} = \langle (x_i - \mu_i)(x_j - \mu_j) \rangle.$$

Possiamo quindi definire una matrice di correlazione come

$$\rho_{ij} = \frac{C_{ij}}{\sqrt{C_{ii}C_{jj}}}.$$

Poi, anche in questo caso di n variabili possiamo definirne l'indipendenza fattorizzando come prima:

$$\rho(x_1, \dots, x_n) = \prod_{i=1}^n \rho_i(x_i).$$

Nel caso specifico di variabili i.i.d., otteniamo $\rho(x_1, \dots, x_n) = \prod_{i=1}^n \rho(x_i)$. Le matrici di covarianza saranno rispettivamente $C_{ij} = \delta_{ij}\sigma_i^2$, e $C_{ij} = \delta_{ij}\sigma^2$ per variabili i.i.d..



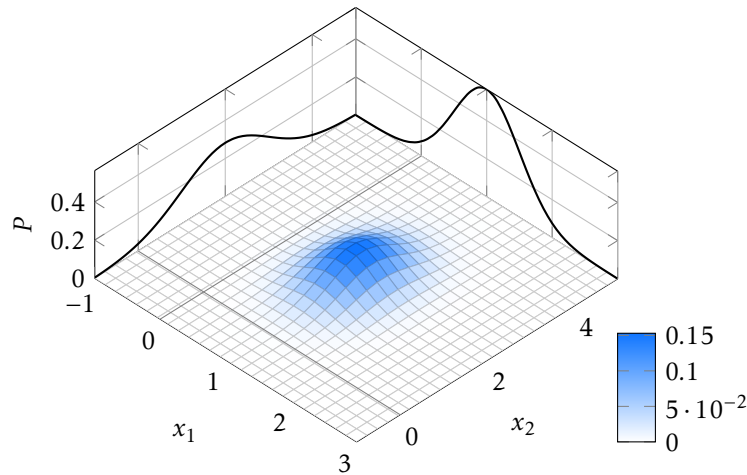


Figura 18: Rappresentazione di una bigaussiana di variabili scorrelate.

Facciamo ancora un esempio con una distribuzione gaussiana, questa volta con variabili correlate dal parametro $\rho \neq 0$. In generale la distribuzione si scrive come

$$\mathcal{N}(x, y) = \frac{1}{\sqrt{(1-\rho^2)2\pi\sigma_x\sigma_y}} \exp\left[-\frac{1}{2(1-\rho^2)}\left(\frac{x^2}{s_x^2} + \frac{y^2}{s_y^2} - \frac{2\rho xy}{s_x s_y}\right)\right].$$

In particolare osserviamo che

$$\langle x^2 \rangle = s_x^2, \quad \langle y^2 \rangle = s_y^2, \quad \langle xy \rangle = \rho s_x s_y$$

e che quindi il coefficiente di correlazione è dato da

$$\rho = \frac{\langle xy \rangle}{\sqrt{\langle x^2 \rangle \langle y^2 \rangle}} = \frac{\rho s_x s_y}{s_x s_y} = \rho.$$

Invece

$$N_x(x) = \frac{1}{\sqrt{2\pi}s_x} e^{-\frac{x^2}{2s_x^2}}, \quad N_y(y) = \frac{1}{\sqrt{2\pi}s_y} e^{-\frac{y^2}{2s_y^2}}.$$

Per $\rho = 0$ le variabili sono scorrelate, e quindi $N(x, y) = N(x)N(y)$, fattorizza.

Come si può caratterizzare questa distribuzione? Una cosa che possiamo fare è calcolare le “curve equiprobabili”, delle curve di livello della PDF. Ci dobbiamo immaginare delle funzioni del tipo

$$\frac{1}{1-\rho^2} \left(\frac{x^2}{s_x^2} + \frac{y^2}{s_y^2} - \frac{2\rho xy}{s_x s_y} \right) = \text{const},$$

equazione di un’ellisse centrata nell’origine. Nel caso estremo $|\rho| = 1$ le variabili sono correlate linearmente, l’ellisse degenera in una retta.

Possiamo trovare l’espressione per l’angolo φ dell’ellisse a partire dalle trasformazioni $x' = x \cos \varphi + y \sin \varphi$ e $y' = -x \sin \varphi + y \cos \varphi$:

$$\tan \varphi = \frac{2\rho s_x s_y}{s_x^2 - s_y^2}.$$

Quando $\varphi = 0$, cioè se l’ellisse è orientata lungo gli assi o è un cerchio, allora le variabili sono scorrelate—cosa molto utile da sapere durante un’analisi dei dati: plottando questi e osservandone la distribuzione sul piano xy , l’impatto a prima vista può già dire qualcosa circa la correlazione, Fig. 19. Invece, quando $\varphi = \pi$, le variabili sono anticorrelate. In conclusione, **quando abbiamo delle distribuzioni approssimabili come gaussiane, studiandone le curve di livello, capiamo che se queste sono orientate in maniera non parallela agli assi ci troviamo in presenza di correlazioni.**

Concludiamo chiedendoci quale sia l’espressione per una gaussiana n -dimensionale:

$$\mathcal{N}(x_1, \dots, x_n) = \frac{1}{\sqrt{(2\pi)^n \det C}} \exp\left[-\frac{1}{2}(\vec{x} - \vec{x}_0) C^{-1} (\vec{x} - \vec{x}_0)\right] = \frac{1}{\sqrt{(2\pi)^n \det C}} \exp\left[-\frac{1}{2} \sum_{ij} (x_i - x_i^0) C_{ij}^{-1} (x_j - x_j^0)\right],$$



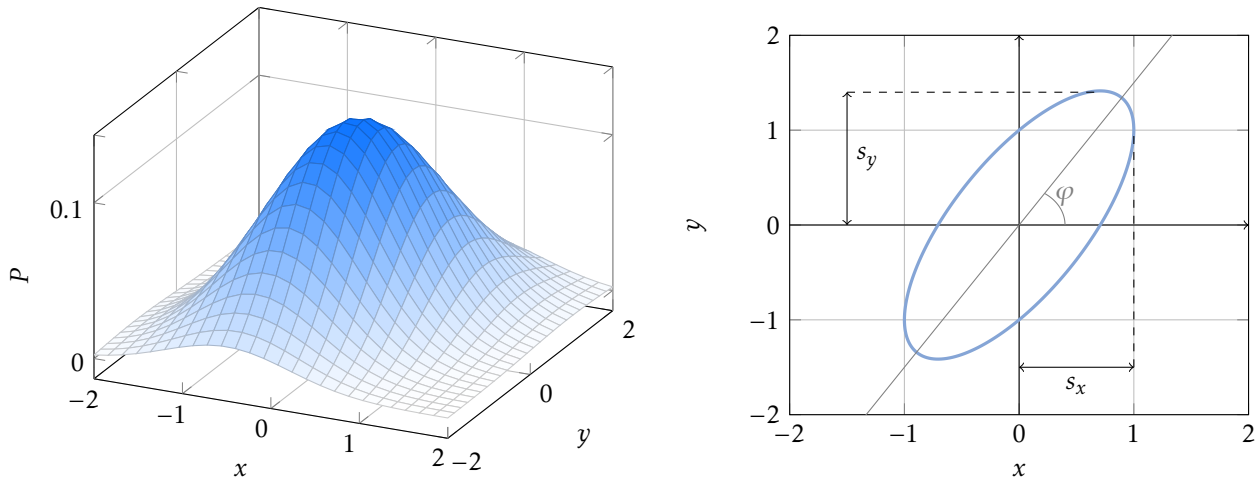


Figura 19: PDF e la corrispondente curva di livello di due variabili gaussiane correlate con $\rho = 1/2$, $s_x = 1$ e $s_y = 1.4$. Tale correlazione tra le variabili ruota l'asse dell'ellisse (in grigio) di un angolo $\varphi = \arctan(1.46) \approx 50^\circ$.

dove $C_{ij} = s_i^2 \delta_{ij} + (1 - \delta_{ij}) \rho_{ij} s_i s_j$ è la matrice

$$C = \begin{pmatrix} s_1^2 & \rho_{12} s_1 s_2 & \cdots & \rho_{1n} s_1 s_n \\ \rho_{21} s_2 s_1 & s_2^2 & & \vdots \\ \vdots & & \ddots & \\ \rho_{n1} s_n s_1 & \cdots & & s_n^2 \end{pmatrix}$$

Definiamo la matrice $V = C^{-1}$, che nel caso bidimensionale di due variabili è:

$$V = \frac{1}{1 - \rho^2} \begin{pmatrix} \frac{1}{s_x^2} & -\frac{\rho}{s_x s_y} \\ -\frac{\rho}{s_x s_y} & \frac{1}{s_y^2} \end{pmatrix}.$$

Adesso andiamo a parlare di distribuzioni estremali o teoria dei valori estremi, utili quando si parla di cercare la media con cui capitano eventi particolari come massimi o minimi. Viene studiato asintoticamente l'andamento di valori estremi di una distribuzione di numeri casuali. Questa teoria è importante in diversi campi, tra cui soprattutto quello della teoria dei materiali e delle fratture.

La resistenza dei materiali (alla frattura) può essere modellizzata in maniera naïve come un ascensore attaccato ad un cavo. Qual è il carico massimo che si può tenere prima che si rompa il cavo? Come faccio a stabilirlo?

In generale, se abbiamo un oggetto soggetto a un campo di deformazione (tipo forza peso), tipicamente questo si deformerà passando da una lunghezza l ad una Δl . Se la deformazione è irreversibile non si torna alla lunghezza a riposo una volta tolto lo sforzo, come ad esempio con il fil di ferro. Per le deformazioni elastiche si torna alla situazione di partenza. La frattura è un caso diverso: può coesistere con la deformazione elastica (materiali duttili).

Il problema di capire come/perché i materiali si rompono ha radici antiche (esperimento di Leonardo). Quello che conta non è il peso ma lo stress, il peso sulla sezione. La resistenza è proporzionale alla sezione, mentre lo sforzo ne è indipendente $\sigma = F/S$ (forza per unità di sezione). In prima approssimazione, la deformazione relativa è pari a

$$\Delta l/l = F/SE,$$

dove E è il modulo di Young.

Possiamo immaginare il fil di ferro come una catena, in cui ogni anello è diverso l'uno dall'altro. In particolare, dove si romperà la catena? Non nel punto medio, ma nel punto di resistenza più debole, e più è lunga la catena e maggiore è la probabilità di trovare una zona con molti difetti. Se conosco la distribuzione dei carichi di rottura della catena, il carico di rottura complessivo sarà dato da quello minimo (dove si rompe). Arriviamo quindi al perché si studia la probabilità degli estremi di una distribuzione.



5 Distribuzioni estremali

5.1 Massimi

Il problema è abbastanza semplice da formulare: consideriamo un sistema di variabili x_1, \dots, x_N i.i.d. distribuite con una certa $p(x_i)$. Sia

$$M = \max_{i=1, \dots, N} (x_1, \dots, x_N).$$

Per ogni variabile possiamo calcolare la funzione cumulata $F(x) = p(\{x_i < x\})$. Per N variabili, avremo

$$F_N(M) = \prod_i p(\{x_i < M\}) = F(M)^N.$$

La PDF sarà data dalla derivata della cumulata (definizione⁻¹ di cumulata...):

$$p_N(M) = NF(M)^{N-1},$$

e questa è la statistica dei massimi.

Prendiamo un primo esempio con la distribuzione uniforme $p(x) = 1$ per $x \in (0, 1)$. Quindi $F(x) = x$, e $F_N(M) = M^N$ —questo è logico: se prendiamo come massimo $M = 1$ abbiamo la certezza che qualunque variabile estratta da una PDF uniforme sarà sotto (o al più) tale valore. Ma allora $p_N(M) = NM^{N-1}$ e quindi

$$\langle M \rangle = \int_0^1 dM M^N N = \frac{N}{N+1} = \frac{1}{1+1/N}.$$

Si osservi come asintoticamente il valor medio del massimo sia 1 per una distribuzione uniforme tra 0 e 1.

Possiamo fare un secondo esempio con la distribuzione esponenziale $p(x) = \lambda e^{-\lambda x}$, con $x > 0$. In questo caso $F(x) = \int_0^x dx' \lambda e^{-\lambda x'} = (1 - e^{-\lambda x})$. Ma allora $F_N(M) = (1 - e^{-\lambda M})^N$ e quindi $p_N(M) = N(1 - e^{-\lambda M})^{N-1} \lambda e^{-\lambda M}$. Segue che

$$F_N(M) = \exp(N \ln(1 - e^{-\lambda M})) \approx \exp(-Ne^{\lambda M}) = \exp(-e^{-\lambda(M - \frac{\ln N}{\lambda})}).$$

Riscriviamola in funzione di due parametri α e β :

$$F(x|\alpha, \beta) = \exp(-e^{-\alpha(x-\beta)})$$

questa distribuzione prende il nome di distribuzione di Gumbel—cioè dei massimi di una distribuzione esponenziale, Fig. 20. Tutte le distribuzioni che hanno un andamento delle code esponenziale (non a potenza) sono asintoticamente uguali alla distribuzione di Gumbel (una specie di CLT). Abbiamo che

$$\langle M \rangle = \frac{\ln N}{\lambda} + \frac{\ln 2}{\lambda} \gamma,$$

non satura mai! Prendendo infiniti campioni, il massimo divergerà logaritmicamente (cfr. Fig. 20).

5.2 Minimi

Ci chiediamo ora se valga un discorso analogo, ma per i minimi di una distribuzione. Sia

$$m = \min_{i=1, \dots, N} \{x_1, \dots, x_N\}.$$

Prendiamo questa volta non la cumulata ma la funzione di sopravvivenza $S(x) = 1 - F(x) = p(\{x' > x\})$. Avremo quindi che, per N variabili,

$$S_N(m) = S(m)^N.$$

Facciamo un esempio con una distribuzione $p(x) = (\alpha + 1)x^\alpha$ per $x \in (0, 1)$. Questa distribuzione è abbastanza generale, dal momento che raggruppa tante distribuzioni diverse al variare del parametro α . Necessariamente si ha che $p(x \rightarrow 0) \rightarrow 0$, e quindi

$$S(x) = \int_x^1 dx' (\alpha + 1)x'^\alpha = 1 - x^{\alpha+1}.$$

Per N variabili, la funzione di sopravvivenza sarà

$$S_N(m) = (1 - m^{\alpha+1})^N = e^{N \ln(1 - m^{\alpha+1})} \approx e^{-Nm^{\alpha+1}} = e^{-\left(\frac{m}{m_0}\right)^{\alpha+1}},$$



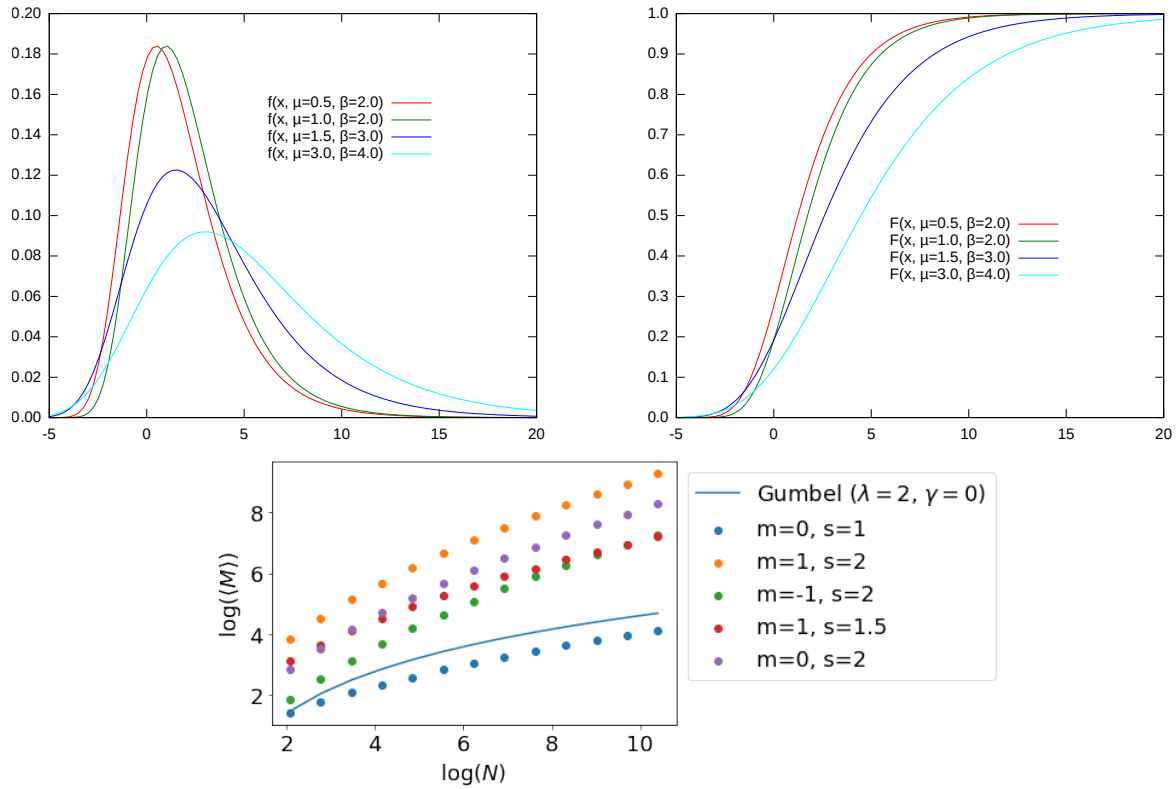


Figura 20: Sopra, PDF e rispettive cumulative della distribuzione di Gumbel, dove $\mu = \alpha$. Sotto, simulazioni della distribuzione di $\langle M \rangle$ presi i massimi da distribuzioni gaussiane di media e varianza variabili, confrontate con il valore teorico $\langle M \rangle = \frac{1}{2} \ln N$.

dove il parametro

$$\frac{1}{m_0^{\alpha+1}} = N \implies m_0 = \left(\frac{1}{N} \right)^{\frac{1}{\alpha+1}}$$

La distribuzione prende il nome di distribuzione di Weibull (Fig. 21), e diventa quindi

$$S_N(x) = e^{-\left(\frac{x}{x_0}\right)^k},$$

per cui si ha che

$$\langle m \rangle \sim N^{-\frac{1}{\alpha+1}}.$$

Riassumendo, se le variabili di partenza sono positive con una distribuzione cumulata che per $x \rightarrow 0$ scala come $P(x) \sim x^a$, il minimo m di N variabili converge alla distribuzione di Weibull per $N \rightarrow \infty$. Ci aspettiamo che il valor medio decresca con N come appena indicato.

5.3 Teorema di Fisher-Tippet-Gnedenko

Enunciamo ora il teorema di Fisher-Tippet-Gnedenko: dati N numeri casuali i.i.d.,

$$\exists a_N, b_N \quad \text{t.c.} \quad \lim_{N \rightarrow \infty} F_N(a_N M + b_N) = G(M) \quad \text{limite invariante,}$$

in cui la funzione $G(M)$ può avere diverse forme:

$$G(M) = \begin{cases} \text{Gumbel} \\ \text{Weibull} \\ \text{Fréchet} \end{cases}$$

abbiamo un **teorema limite simile al CLT**, per cui **esiste una forma asintotica per i massimi**. Il **limite dipende dall'andamento (asintotico) della distribuzione di partenza**, un po' come nel CLT in cui media e varianza dipendevano dalla distribuzione di partenza.



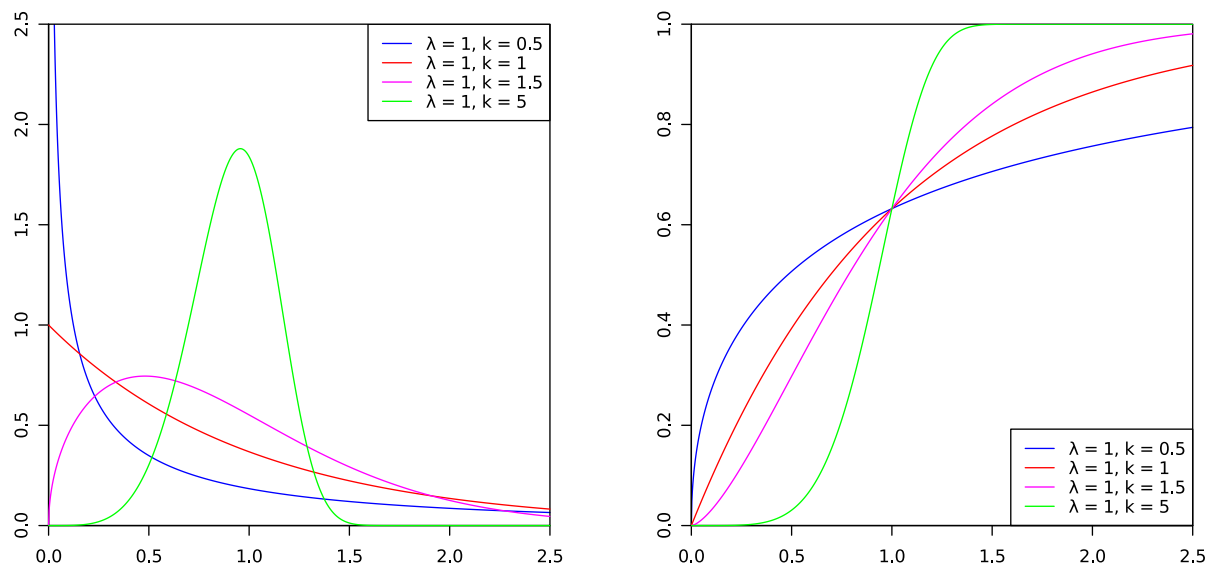


Figura 21: PDF e rispettive cumulative della distribuzione di Weibull.

5.4 I boxplots

Facciamo una parentesi sui boxplots, molto usati con il pacchetto `seaborn` su `python`. I boxplots sono un metodo per rappresentare graficamente dei dati attraverso i loro quartili. I quartili sono quei valori/modalità che ripartiscono la popolazione in quattro parti di uguale numerosità. Importante è il secondo quartile, che coincide con la mediana e divide la popolazione in due parti di uguale numerosità, delle quali il primo ed il terzo quartile sono le mediane. Il quartile zero coincide con il valore minimo della distribuzione. Il quarto quartile coincide con il valore massimo della distribuzione.

Un boxplot è un modo standardizzato di visualizzare un dataset basandosi su cinque valori: il minimo, il massimo, la mediana del campione, il primo e il terzo quantile. Il minimo è ovviamente il dato di valore più basso escludendo gli outliers. Il maggiore è definito in maniera analoga. La mediana è il valore che divide la popolazione in due metà. Il primo quartile Q_1 è la mediana della metà inferiore del dataset (il 25% dei dati sta sotto questo valore), mentre il terzo quartile Q_3 è la mediana della metà superiore (il 75% dei dati sta sotto questo valore). Il range interquartile è definito da $IQR = Q_3 - Q_1$. Solitamente questo range viene utilizzato per definire gli outliers: se un punto si trova sotto $Q_1 - 1.5 \times IQR$ o oltre $Q_3 + 1.5 \times IQR$, viene considerato outlier.

Gli interquartili sono una sorta di deviazione standard, ma sono “più resistenti” di questi alla presenza di outliers. Si consideri l'esempio di due insiemi di dati $\{1, 1, 1, 1, 1, 1, 1\}$ e $\{1, 1, 1, 1, 1, 1, 100000000\}$. Per entrambi i sistemi IQR è nullo, mentre la deviazione standard è molto diversa. L' IQR indica dove si trova il 50% dei dati, mentre la deviazione standard indica quanto questi dati siano sparsi. Tuttavia, la deviazione standard è un righello per una distribuzione normale che ci permette di capire quanto spesso/sia probabile che accada un determinato evento (regola del “68-95-99.7%” in termini della SD). Un esempio comodo per capire l'utilizzo della deviazione standard è il seguente: considerare due ipotetiche offerte di lavoro: la prima è nella città di campagna A, in cui i salari medi sono distribuiti secondo una normale $\mathcal{N}(30k, 5k)$; l'altra è una metropoli B in cui i salari sono distribuiti secondo $\mathcal{N}(60k, 10k)$. Si supponga che ci sia un'offerta per la città A di 45k, mentre l'altra per la città B di 60k. Quale delle due conviene scegliere? La risposta è (ovviamente) la prima, perché l'offerta di 45k dista 3σ (in positivo!) rispetto al valor medio degli stipendi nella città A. Invece, il secondo stipendio, sebbene sia superiore, è conforme alla distribuzione degli stipendi nella città B.

Gli outliers sono valori che differiscono in maniera significativa dagli altri valori osservati/misurati. Un outlier può essere dovuto alla variabilità nella misurazione o può indicare un errore sperimentale; nell'ultimo caso si tende solitamente a escludere questi valori dal dataset. Spesso questi valori indicano degli errori sperimentali di misura o segnalano la presenza di code grasse, non trascurabili, della distribuzione. Nel primo caso si tende a scartare questi valori, mentre nel secondo indicano che la skewness della distribuzione è elevata e quindi bisogna fare attenzione ad assumere distribuzioni normali. Nel caso di dati distribuiti normalmente, la regola delle tre sigma dice che circa 1 caso ogni 22 differisce per il doppio della deviazione standard o più rispetto la media, e che 1 su 370 differisce per tre deviazioni standard. Se su un campione di 1000 dati ci sono cinque osservazioni distanti più di tre sigma è tutto normale, mentre se la taglia del campione è di 100 valori,



tre outliers dovrebbero essere motivo di preoccupazione, essendo più di 11 volte del valore atteso di outliers.

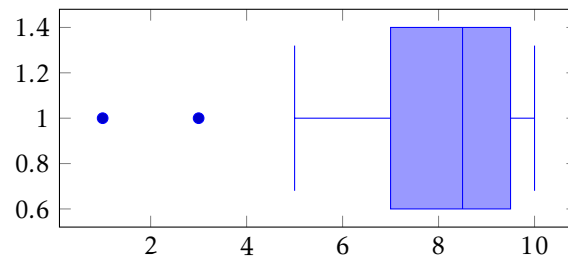


Figura 22: Esempio di un boxplot, con rappresentati alcuni outliers come \bullet .

6 Inferenza statistica

Oggi parliamo di inferenza statistica, cioè della possibilità di valutare se una certa ipotesi è corretta o di trovarne una migliore. Abbiamo dei dati e vogliamo fare ipotesi su questi, per vedere se dicono qualcosa o no. Generalmente sono incompleti e sporcati da rumore esterno, ma ci chiediamo comunque quale sia la realtà che maggiormente si collega a questi.

Ad esempio, prendiamo un mazzo di carte, e supponiamo di tirarne fuori otto ($J\heartsuit, Q\heartsuit, Q\clubsuit, K\spadesuit, K\heartsuit, 8\clubsuit, 9\heartsuit$ e $7\spadesuit$): da che mazzo sono state tirate fuori? Abbiamo due possibilità: (i) sono state tirate fuori da un mazzo francese completo di 54 carte, (ii) oppure sono state tirate fuori da un mazzo da poker con 32 carte. Qual è l'ipotesi più pertinente? Possiamo stabilire quale sia la probabilità che esca fuori con un'estrazione questo insieme di carte, e valutare la probabilità maggiore. Chiaramente, in un mazzo di 32 carte la probabilità che non esca nessuna carta < 6 è 1: e quindi la likelihood $L = 1$, mentre nel caso (i) è $L = (8/13)^8 \approx 0.02 \ll 1$. In questo caso, statisticamente diremmo che è più verosimile che il mazzo sia di 32 carte. Avevamo discusso della probabilità Bayesiana, e anche in questo caso si può disquisire sul fatto che abbiamo dato la stessa probabilità a priori che questo sia un mazzo da poker o un mazzo francese. Ma se fossimo a casa di un giocatore di bridge potremmo dare un bias diverso... abbiamo dato lo stesso peso alle due ipotesi.

Questa è una cosa importante, perchè anche nella vita pratica ci vengono date delle informazioni statistiche, e viene sempre dato a intendere che il campione sia unbiased e che le ipotesi siano tutti equamente ragionevoli. Bisogna sempre pensare però quando si leggono dei dati statistici quali siano le ipotesi e quale sia il contesto.

Un altro possibile esempio è quello di avere un'università in cui in un corso di laurea di siano 532 maschi e 490 femmine. Questa distribuzione è compatibile con una distribuzione $N_M = N_F$? Dobbiamo calcolare la probabilità che avvenga questo esempio e valutarne la likelihood: 523 è tanto distante da 490 o no? Dipende appunto dal contesto in cui valutiamo questi dati. La vera ipotesi è che la probabilità che si iscriva un maschio sia uguale a quella in cui si iscriva una femmina $P_M = P_F$. Posso immaginare che queste fluttuazioni siano dovute al caso oppure è un effetto vero?

Un altro esempio ancora è quello in cui abbiamo una serie di particelle con tempo di vita diverso τ_1, \dots, τ_N . Quanto vale τ ? Quanto è la sua incertezza? Si tratta sempre dello stesso problema: cerchiamo di inferire una realizzazione della realtà di fondo, che soffre di problemi come errori di misura, fluttuazioni termiche etc..

Proviamo a vedere questi concetti in modo più formale. Immaginiamo di avere un certo dato: consideriamo N possibili ipotesi alternative H_i . Qual è la probabilità che valga H_i , che valga ciascuna delle ipotesi? Un caso più semplice è quello in cui abbiamo una sola ipotesi H_0 , e cerchiamo a questo punto di stabilire se H_0 è vera o falsa. Un terzo caso ancora è quello in cui l'ipotesi dipende con continuità da un (o più) parametro $H = H(\lambda)$.

Queste tre formalizzazioni si collegano con i tre esempi fatti prima. Cominciamo a vedere come possiamo in generale affrontare questo problema, con un approccio moderatamente bayesiano. Partiamo dall'esempio delle ipotesi discrete (primi due esempi): immaginiamo di osservare un certo esempio k . Abbiamo poi delle ipotesi (che abbiamo detto essere discrete) H_1, \dots, H_N . Per stabilire quale sia più verosimile dobbiamo calcolare la probabilità che valga $p(H_i|k)$, la probabilità che valga H_i dato k . Sappiamo che possiamo facilmente conoscere la probabilità inversa, cioè che dato H_i valga k : $p(k|H_i)$. Possiamo usare il teorema di Bayes:

$$p(H_i|k) = \frac{p(k|H_i)p(H_i)}{p(k)}.$$

Possiamo scrivere $p(k) = \sum_i p(k|H_i)p(H_i)$, per cui si trova il risultato

$$p(H_i|k) = \frac{p(k|H_i)p(H_i)}{\sum_i p(k|H_i)p(H_i)}.$$



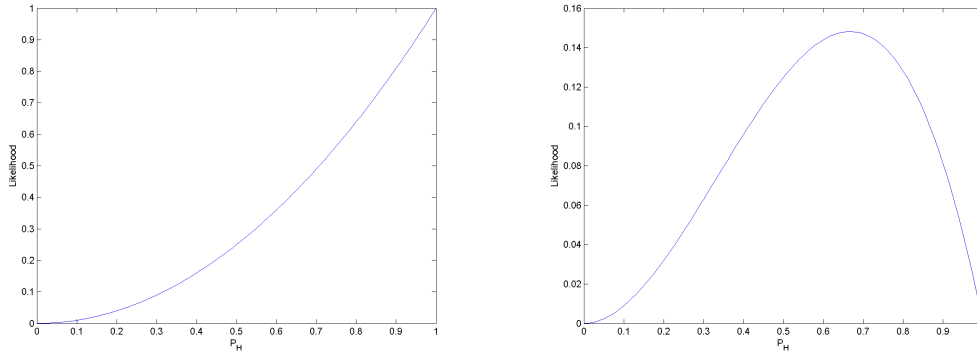


Figura 23: Confronto tra likelihoods. A sinistra, la likelihood per la probabilità p_T^2 che una moneta tirata due volte faccia testa-testa (senza conoscenza a priori sul fatto che sia truccata o meno). A destra, la likelihood per la probabilità $p_T^2(1 - p_T)$ per un'osservazione testa-testa-croce, sempre senza prior.

Quello che non conosciamo è la probabilità che sia valida l'ipotesi. Questo è il classico esempio di prior, di assunzione che facciamo: dipende dal contesto la nostra ipotesi.

Spesso ci basta la probabilità relativa, per cui ci basterebbe calcolare il rapporto tra queste probabilità:

$$\frac{p(H_i|k)}{p(H_j|k)} = \frac{p(k|H_i)p(H_i)}{p(k|H_j)p(H_j)} = \frac{p(k|H_i)}{p(k|H_j)}.$$

Facciamo adesso un altro esempio: consideriamo una moneta, per cui abbiamo due possibili ipotesi: H moneta truccata, e \bar{H} moneta non truccata. L'evento k è che siano uscite due teste dopo due tiri. In questo caso $p(k|H) = 1$ (se la moneta è truccata, allora è "normale" che dopo due tiri escano due teste), e $p(k|\bar{H}) = 1/2 \times 1/2 = 1/4$ (se non è truccata faccio il calcolo con le probabilità standard che escano due teste dopo due tiri). Quindi

$$\frac{p(H|k)}{p(\bar{H}|k)} = 4,$$

è quattro volte più probabile che la moneta sia truccata. Oppure, in altri termini,

$$p(H|k) = \frac{1}{1 + \frac{1}{4}} = \frac{4}{5} \implies p(\bar{H}|k) = \frac{1}{5}.$$

Tutto questo vale con un prior uniforme, piatto: è ragionevole pensare che la moneta non sia truccata, ma è più ragionevole pensare che lo sia. Ricordiamoci che questo vale perché abbiamo fatto due tiri soli: aumentando i tiri, a seconda degli esiti il profilo della likelihood cambia (cfr. Fig. 23).

Facciamo un altro passo in avanti, e consideriamo inferenza continua, con dipendenza da un parametro. Abbiamo una certa osservazione x e un'ipotesi $H = H(\theta)$. Ci chiediamo quale sia la probabilità che valga θ dato il risultato x . Come al solito utilizziamo il teorema di Bayes (Fig. 2):

$$p_x(x|\theta)\Pi_\theta(\theta) = p_\theta(\theta|x)\Pi_x(x) \implies p_\theta(\theta|x) = \frac{p_x(x|\theta)\Pi_\theta(\theta)}{\Pi_x(x)}.$$

Imponendo la marginalizzazione

$$\Pi_x(x) = \int_{-\infty}^{\infty} d\theta p_x(x|\theta)\Pi_\theta(\theta),$$

otteniamo che

$$p_\theta(\theta|x) = \frac{p_x(x|\theta)\Pi_\theta(\theta)}{\int_{-\infty}^{\infty} d\theta p_x(x|\theta)\Pi_\theta(\theta)}.$$

Cercando il massimo rispetto a θ troviamo l'ipotesi più probabile: questo è il principio di massima verosimiglianza. Tipicamente, quando vogliamo massimizzare questa funzione usiamo un prior uniforme, in cui $\Pi_\theta(\theta)$ su tutti i possibili parametri. In questo caso $p_\theta(\theta|x) = p_x(x|\theta)$, opportunamente normalizzato. Avendo più variabili $\{x_1, \dots, x_N\}$ abbiamo

$$L(\theta) = \prod_{i=1}^N p(x_i|\theta)$$



come funzione di verosimiglianza. Questa funzione fa il prodotto di tutte le probabilità che esca x data la validità di θ .

Arriviamo quindi al **Principio di Massima Verosimiglianza**, che ci dà la ricetta per trovare il θ più probabile, dato da $\max_{\theta} L(\theta)$. Spesso è più semplice lavorare con il logaritmo della likelihood, dal momento che se nella likelihood abbiamo il prodotto delle probabilità, nella log-likelihood si ottiene la somma dei loro logaritmi:

$$L' = \ln L(\theta) = \sum_{i=1}^N \ln p(x_i|\theta),$$

il cui massimo $\hat{\theta}$ è dato dall'equazione

$$\left. \frac{d \ln L}{d \theta} \right|_{\hat{\theta}} = 0.$$

Il criterio di massima verosimiglianza ci dice che la configurazione più probabile si trovava massimizzando la (log-)likelihood rispetto al parametro θ . Per fare un esempio, consideriamo una serie di variabili casuali gaussiane x_i , con $i = 1, \dots, N$. Ci viene data la varianza σ , e vogliamo stimare il valore di μ che meglio si adatta a questi dati. Il nostro modello interpretativo

$$p(x|\mu) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

Dobbiamo quindi calcolare $L(\mu) = \prod_i p(x_i|\mu)$, o più semplicemente il suo logaritmo

$$\ln L = - \sum_i \frac{(x_i - \mu)^2}{2\sigma^2} + c = - \frac{1}{2\sigma^2} \sum_i (x_i^2 - 2\mu x_i + \mu^2) + c,$$

per cui

$$\frac{\partial \ln L}{\partial \mu} = - \frac{1}{2\sigma^2} (-2\bar{x}N + 2\mu), \quad \bar{x} = \frac{1}{N} \sum_i x_i.$$

Possiamo scrivere a questo punto

$$0 \stackrel{!}{=} \frac{\partial \ln L}{\partial \mu} = \frac{1}{\sigma^2} (\bar{x} - \mu) \implies \mu = \bar{x},$$

giustamente esce la media, *anche* dal principio di massima verosimiglianza. La media è la stima migliore per il parametro μ della gaussiana.

Facciamo un altro esempio, immaginando di avere x_i gaussiane, dove questa volta conosciamo μ e vogliamo stimare σ come

$$p(x|\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

In questo caso

$$\ln L(\sigma) = - \sum_i \frac{(x_i - \mu)^2}{2\sigma^2} - N \ln \sigma + c.$$

Quindi

$$0 \stackrel{!}{=} \frac{\partial \ln L}{\partial \sigma} = \frac{1}{\sigma^3} \sum_i (x_i - \mu)^2 - \frac{N}{\sigma},$$

da cui

$$\frac{1}{\sigma^2} \sum_i (x_i - \mu)^2 = N \implies \sigma^2 = \sum_i \frac{(x_i - \mu)^2}{N}.$$

Tutto questo si può generalizzare in presenza di più parametri. Uno può definire una funzione di verosimiglianza $L(\lambda_1, \dots, \lambda_k) = \prod_{i=1}^N p(x_i|\lambda_1, \dots, \lambda_k)$ per cui

$$\ln L = \sum_i \ln p(x_i|\lambda_1, \dots, \lambda_k).$$

L'estremo va trovato rispetto a tutte le i variabili

$$\frac{\partial \ln L}{\partial \lambda_i} = 0$$



per $i = 1, \dots, k$. La funzione di verosimiglianza sarà una funzione di uno spazio multiparametrico, in cui il massimo sarà dato da $\hat{\lambda}_i$, e avremo delle curve di livello della funzione che ci daranno le barre di errore—contrariamente al caso dipendente da un parametro, dove abbiamo la “larghezza” della funzione, un po’ come la varianza nella gaussiana. Facciamo un esempio per capire questo concetto: prendiamo la solita serie di variabili gaussiane $\{x_i\}$, e vogliamo stimare sia μ che σ . Quindi

$$\ln L = - \sum_i \frac{(x_i - \mu)^2}{2\sigma^2} - N \ln \sigma + c,$$

da cui segue che

$$\begin{aligned} \frac{\partial \ln L}{\partial \mu} &\stackrel{!}{=} 0 \implies \mu \sum_i 1 = N\mu \stackrel{!}{=} \sum_i x_i \implies \hat{\mu} = \bar{X} \\ \frac{\partial \ln L}{\partial \sigma} &\stackrel{!}{=} 0 \implies +\frac{1}{\sigma^3} \sum_i (x_i - \mu)^2 - \frac{N}{\sigma} \stackrel{!}{=} 0 \implies \hat{\sigma}^2 = \frac{1}{N} \sum_i (x_i - \mu)^2 = \overline{(x - \bar{x})^2}. \end{aligned}$$

6.1 Minimi quadrati

Abbiamo due variabili, di cui una con un’incertezza x_i e $y_i + \delta_i$. Qual è la relazione tra queste?

$$y = t(x, \vec{\theta}).$$

Dobbiamo calcolare il chi quadro associato a queste variabili:

$$\chi^2 = \sum_{i=1}^N \frac{(y_i - t(x_i, \vec{\theta}))^2}{\delta_i^2},$$

devo pesare rispetto all’incertezza sulla misura. Questo è quello che si fa con il chi quadro. Se immaginiamo che le osservazioni siano gaussiane, possiamo dire che

$$p(y_1, \dots, y_n | \vec{\theta}) \propto \exp \left(- \sum_i \frac{(y_i - t(x_i, \vec{\theta}))^2}{2\delta_i^2} \right)$$

e che quindi

$$\ln L = - \frac{1}{2} \sum_i \frac{(y_i - t(x_i, \vec{\theta}))^2}{\delta_i^2}.$$

Abbiamo quindi che

$$\ln L = - \frac{1}{2} \chi^2.$$

Questo vuol dire che **se gli errori sono gaussiani, allora il test del chi quadro** (minimi quadrati) non è altro che il **criterio di massima verosimiglianza**! Quest’ultimo è tuttavia più generico, perché non assume nessuna distribuzione, mentre il test del chi quadro assume la gaussianità delle variabili sottostanti. Volendo, uno può scrivere esplicitamente il valore delle $y_i = ax_i + b$ e $\vec{\theta} = (a, b)$ e calcolare il minimo del chi quadro, trovando esplicitamente i valori di a e b che danno il risultato migliore $t = ax + b$.

6.2 Propagazione degli errori

Come si propagano gli errori? Come possiamo definire degli errori di interessi quando abbiamo una dipendenza di una variabile da un’altra variabile? Abbiamo $x \pm \delta x$, dove $\delta x = \sqrt{\langle \Delta x^2 \rangle}$, e Δx è lo scostamento dal valor medio. Se abbiamo una certa funzione $y(x)$, come associamo a questa un errore δy ? Assumendo il valor medio x_m possiamo scrivere

$$y(x_m \pm \delta x) \approx y(x_m) \pm \frac{\partial y}{\partial x} \delta x,$$

da cui

$$\langle \delta y^2 \rangle = \langle (y - y_m)^2 \rangle \approx \left| \frac{\partial y}{\partial x} \right|^2 \delta x^2 \implies \delta y = \left| \frac{\partial y}{\partial x} \right| \delta x.$$



Nel caso di due variabili $y(x_1, x_2)$ con incertezze δx_1 e δx_2 , avremo

$$y(x_1 \pm \delta x_1, x_2 \pm \delta x_2) \approx y(x_1, x_2) \pm \frac{\partial y}{\partial x_1} \delta x_1 \pm \frac{\partial y}{\partial x_2} \delta x_2,$$

da cui segue che

$$\langle \delta y^2 \rangle = \left(\frac{\partial y}{\partial x_1} \delta x_1 + \frac{\partial y}{\partial x_2} \delta x_2 \right)^2.$$

Sviluppando il quadrato si ottiene che

$$\delta y^2 = \left(\frac{\partial y}{\partial x_1} \right)^2 \delta x_1^2 + \left(\frac{\partial y}{\partial x_2} \right)^2 \delta x_2^2 + 2 \left(\frac{\partial y}{\partial x_1} \right) \left(\frac{\partial y}{\partial x_2} \right) R_{12} \delta x_1 \delta x_2.$$

Interessandoci il valor medio,

$$\langle \delta y^2 \rangle = \left(\frac{\partial y}{\partial x_1} \right)^2 \delta x_1^2 + \left(\frac{\partial y}{\partial x_2} \right)^2 \delta x_2^2 + 2 \left(\frac{\partial y}{\partial x_1} \right) \left(\frac{\partial y}{\partial x_2} \right) \langle \Delta x_1 \Delta x_2 \rangle.$$

Definiamo di conseguenza

$$R_{12} = \frac{\langle \Delta x_1 \Delta x_2 \rangle}{\delta x_1 \delta x_2}$$

coefficiente di cross-correlazione di cui bisogna tener conto. Se le variabili sono scorrelate $R_{12} = 0$.

Possiamo generalizzare questo discorso a un numero indefinito di variabili:

$$\delta y^2 = \sum_{i,j} \frac{\partial y}{\partial x_i} \frac{\partial y}{\partial x_j} R_{ij} \delta x_i \delta x_j, \quad R_{ij} = \frac{\langle \Delta x_i \Delta x_j \rangle}{\delta x_i \delta x_j},$$

tale che $R_{11} = 1$ e $R_{ij} = R_{ji}$. La matrice di covarianza diventa

$$C_{ij} = R_{ij} \delta x_i \delta x_j$$

e quindi

$$\delta y^2 = \sum_{i,j} C_{ij} \frac{\partial y}{\partial x_i} \frac{\partial y}{\partial x_j}.$$

Fino ad adesso abbiamo parlato di errori statistici. Ovviamente, questo non è il solo tipo di errore. Quello più insidioso è quello sistematico: possono esserci molti casi in cui l'errore non ha a che fare con la statistica (ripetizione della misura) ma dal fatto che ci sono fattori esterni che influenzano la misura. **Non esiste soluzione universale per gli errori sistematici.**

Esercizio: date due misure $(x_1, y_1 \pm \delta y_1)$ e $(x_2, y_2 \pm \delta y_2)$ della retta $y = mx + b$, calcolarne l'incertezza.

Le funzioni su cui dobbiamo propagare gli errori sono quella del coefficiente angolare e dell'intercetta, ovvero ci servono l'incertezza sul coefficiente angolare

$$m = (y_2 - y_1)/(x_2 - x_1)$$

e quella sull'intercetta. Quest'ultima si ottiene sostituendo l'espressione per il coefficiente angolare dentro una delle due equazioni della retta in forma implicita:

$$\begin{aligned} y_1 &= mx_1 + b \implies b = y_1 - mx_1 \\ &= y_1 - \frac{y_2 - y_1}{x_2 - x_1} x_1 \\ &= \frac{y_1(x_2 - x_1) - (y_2 - y_1)x_1}{x_2 - x_1} \\ &= (x_2 y_1 - x_1 y_2)/(x_2 - x_1). \end{aligned}$$

In accordo con

$$\delta y^2 = \sum_{i,j} \frac{\partial y}{\partial x_i} \frac{\partial y}{\partial x_j} R_{ij} \delta x_i \delta x_j, \quad R_{ij} = \frac{\langle \Delta x_i \Delta x_j \rangle}{\delta x_i \delta x_j},$$



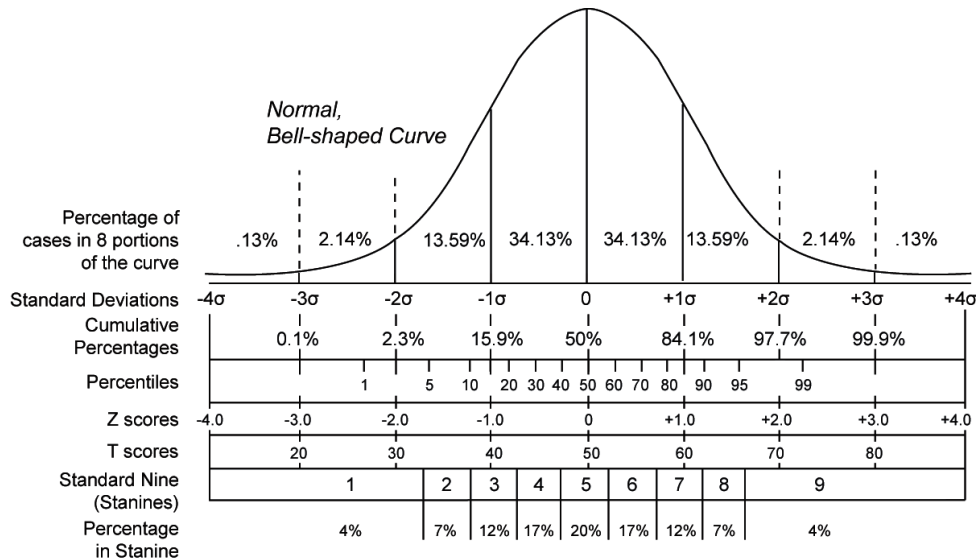


Figura 24: Rappresentazione degli intervalli di confidenza di una distribuzione gaussiana.

otteniamo

$$\delta m^2 = \frac{\delta y_2^2 + \delta y_1^2}{(x_2 - x_1)^2}$$

$$\delta b^2 = \frac{x_2^2 \delta y_1^2 + x_1^2 \delta y_2^2}{(x_2 - x_1)^2}$$

$$\langle \Delta m \Delta b \rangle = -\frac{x_2 \delta y_1^2 + x_1 \delta y_2^2}{(x_2 - x_1)^2}.$$

6.3 Intervalli di confidenza

Tornando alla questione degli intervalli di confidenza, che cosa è tale intervallo? È una misura di “quanto noi ci fidiamo di questa misura”, e tipicamente gli errori seguono una distribuzione gaussiana. Se la distribuzione è gaussiana, facciamo una misura x e ne associamo un certo errore δ . Qual è il livello di confidenza di questa misura? Prendendo come errore la σ , vogliamo calcolare

$$I(\delta) = \int_{-\delta}^{\delta} \frac{dx}{\sqrt{2\pi}\sigma} e^{-\frac{(x-x_m)^2}{2\sigma^2}},$$

probabilità che la misura cada nell’intervallo $[x_m - \delta, x_m + \delta]$. Per esempio, il bosone di Higgs venne annunciato quando il livello di confidenza era 5σ , la probabilità che l’evento fosse in quel regime e non fosse dovuto a rumore era elevatissima. Nella fisica delle particelle ci sono tanti falsi positivi entro 3σ , per cui si va su intervalli molto più grandi.

Se cerchiamo la massima verosimiglianza di $\ln L(\theta)$, il modo per stimare gli intervalli di confidenza è fare un’approssimazione parabolica, dal momento che la log-likelihood di una gaussiana è al primo ordine una parabola (strano da dire...)—e l’intervallo di errore corrisponde alla regione in cui si passa dal valore massimo alla sua metà—, per cui

$$-\ln L(\theta) \approx \frac{1}{2} V(\theta - \hat{\theta})^2 \implies \delta\theta^2 = V^{-1} = -\left(\frac{d^2 \ln L}{d\theta^2} \bigg|_{\hat{\theta}} \right)^{-1}.$$

Quindi, nel caso di dipendenza da più parametri,

$$-\ln L \approx \frac{1}{2} \sum_{i,j} (\theta_i - \hat{\theta}_i) V_{ij} (\theta_j - \hat{\theta}_j),$$



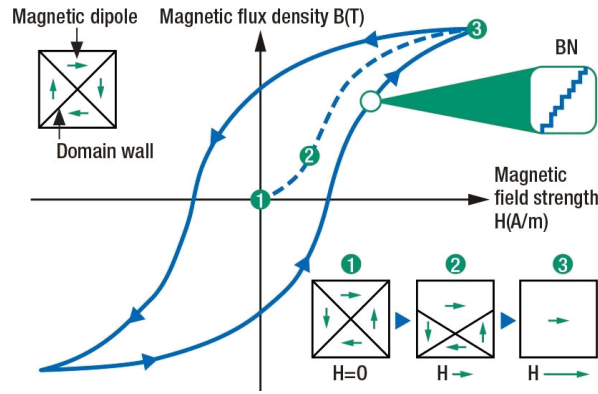


Figura 25: Ciclo di isteresi e rumore di Barkhausen.

dove la matrice

$$V_{ij} = - \left. \frac{\partial^2 \ln L}{\partial \theta_i \partial \theta_j} \right|_{\hat{\theta}}$$

è la matrice dei pesi, ottenuta tramite le varie derivate. Da questa si ottiene la matrice delle correlazioni per inversione $C = V^{-1}$.

6.4 Rumore Barkhausen

Il rumore Barkhausen è un fenomeno abbastanza comune nei ferromagneti, che a campo nullo hanno magnetizzazione non nulla. La proprietà fondamentale dei ferromagneti è quello di avere un ciclo di isteresi: plottando $M = M(H)$ si trova un andamento tipo quello della Fig. 25. L'area di questo "loop" rappresenta l'energia dissipata durante il processo. I sistemi paramagnetici non hanno isteresi. Nel 1919, Barkhausen scoprì che guardando con una certa risoluzione la curva di isteresi non è liscia, smooth ma è a scalino → salti della magnetizzazione, non è regolare così come sembra. I momenti magnetici cambiano a salti. Più precisamente, l'effetto (rumore) Barkhausen è una serie di cambiamenti repentini nella taglia e nell'orientazione dei domini ferromagnetici, cioè nei cluster microscopici di spin allineati.

Questi salti nella magnetizzazione sono causati da cambiamenti discreti nella taglia o rotazione di questi domini ferromagnetici. I domini cambiano taglia per via dei *domain walls* nel cristallo reticolare che si muovono in seguito a dei cambiamenti nel campo magnetico. In un cristallo perfetto questo può essere un processo continuo, ma nei cristalli reali difetti locali nel reticolo (come impurità o dislocazioni nella struttura) causano l'avvolgimento dei domain walls attorno a questi difetti (immaginare tipo un fronte d'onda che supera uno scoglio, richiudendosi dietro di esso). Quando il cambiamento del campo magnetico diventa abbastanza forte da superare la barriera energetica del difetto, vengono flippati contemporaneamente tutti gli spin del dominio "aggrappato" in contemporanea, e il domain wall si "teletrasporta" e supera la barriera. **Questo cambiamento improvviso nella magnetizzazione causa un cambiamento nel flusso magnetico attraverso la sbarretta, che è registrata dall'esterno come un "click" nell'amplificatore**, da cui il nome del fenomeno. Le fluttuazioni riflettevano la dinamica di queste magnetizzazioni.

Il voltaggio che misuriamo è $V \propto dH/dt$, i momenti magnetici positivi cercheranno di aumentare. Questo viene dalle leggi di induzioni delle equazioni di Maxwell. Vogliamo studiare le fluttuazioni di questa velocità di magnetizzazione. Ci deve essere qualcosa che rende il tutto più fluttuante, e questo qualcosa è il disordine presente nei materiali, che ovviamente non sono perfetti ma presentano impurezze che bloccano il moto della parete. Da queste parte un modello matematico che prova a descrivere questo rumore (modello ABBM). Il modello si basa sull'idea di considerare una singola parete magnetica, e consideriamo un solo grado di libertà x . Assumendo $M(x=0) = 0$, troviamo $M = M_S x/L$, dove $2L$ è la lunghezza del campione (per cui $x \in [-L, L]$). Scriviamo un'equazione per il moto viscoso di questa magnetizzazione, con viscosità Γ (che assumiamo poi unitaria):

$$\Gamma \frac{dx}{dt} = H(t) - kx + W(x),$$

dove l'ultimo termine tiene conto delle impurezze e del disordine (parte statistica). $W(x)$ è un campo casuale, che prendiamo gaussiano, con media nulla e con correlazioni di tipo diffusivo nello spazio

$$\langle |W(x) - W(x')|^2 \rangle = D|x - x'|,$$



un po' come un random walk nello spazio. Possiamo provare a risolvere questa equazione per un campo magnetico crescente nel tempo $H(t) = ct$ e ponendo $\Gamma = 1$:

$$\frac{dx}{dt} = ct - kx + W(x).$$

Cerchiamo l'equazione per la velocità derivando rispetto al tempo:

$$\frac{dv}{dt} = c - kv + v\eta(x),$$

dove $\eta(x) := dW/dx$. Ma questa è la derivata di un RW, cioè un rumore bianco \rightarrow scorrelazione, per cui

$$\langle \eta(x)\eta(x') \rangle = D\delta(x - x'),$$

perché il RW è per definizione l'integrale di un rumore bianco. Con questa nuova equazione scriviamo

$$\frac{dv}{dt} = \frac{dv}{dx} \frac{dx}{dt} = \frac{dv}{dx} v,$$

da cui segue che

$$\frac{dv}{dx} = \frac{c}{v} - k + \eta(x).$$

Supponendo che il tempo sia x , otteniamo un'equazione di Langevin, scritta solitamente come

$$\frac{dy}{dt} = -\frac{\partial U}{\partial y} + \eta(t), \quad k_B T = D.$$

L'equazione di Langevin non è altro che l'equazione di Newton con una forza risultante data da quella applicata e da un termine stocastico. Qual è la probabilità di trovare una certa velocità nel limite stazionario? Per un'equazione di Langevin la risposta è

$$p(y) \propto e^{-\beta U(y)},$$

fattore di Boltzmann \rightarrow sfruttiamo questo mapping sul nostro modello di parete, ottenendo quindi

$$U(v) \stackrel{!}{=} kv - c \ln v.$$

Quindi, adottando il pedice $_x$ per ricordare che non abbiamo propriamente un'equazione di Langevin,

$$p_x(v) \propto e^{-\frac{kv + c \ln v}{D}} = v^{c/D} e^{-kv/D}.$$

Nel caso in cui siamo interessati ad un ensemble fisico (con il tempo), si dimostra che

$$p_t(v) = \frac{1}{v} p_x(v),$$

sono molto più probabili casi con velocità bassa (probabilità maggiore)! A questo punto

$$p_t(v) = \frac{v^{\frac{c}{D}-1} e^{-\frac{k}{D}v}}{\Gamma(c/D)} \equiv G\left(v \left| \frac{c}{D}, \frac{k}{D} \right.\right),$$

è una funzione Gamma (Sec. 2.7)!

7 Il principio di massima entropia

Il principio di massima entropia (detto anche MaxEnt) prende spunto da una definizione dell'entropia, quella di Shannon. È un concetto noto alla teoria dell'informazione e descrive la mancanza di questa all'interno di una distribuzione di probabilità. MaxEnt viene utilizzato tipicamente per stabilire qual è la distribuzione che è maggiormente compatibile con quello che conosciamo (conoscenza a priori in termini Bayesiani), ma con il minimo di assunzioni possibili rispetto a tutte le probabilità \rightarrow scelta "meno informata". È un problema essenziale in tutta l'analisi Bayesiana: rimane sempre la distribuzione del prior che non è nota $p(A)$ che non abbiamo



modo esplicito di assumere. Facevamo quindi delle assunzioni, e MaxEnt ci dà un modo più formalizzato di stabilire il prior. Ovviamente tutto è compatibile con ciò che sappiamo (vincoli).

Cominciamo il seguente esempio: abbiamo un dado a sei facce e ci chiediamo il dato migliore perché esca una delle sei facce del dado $\{q_1, \dots, q_6\}$. Assumiamo il vincolo $\mu_0 = \sum_i q_i = 1$, che le probabilità devono essere normalizzate. Diciamo anche che la media sia $\mu_1 = \sum_i i q_i = 3.5$. In realtà con solo queste informazioni ci sono infinite PDF che soddisfano questi requisiti. Ad esempio, la distribuzione $\{\frac{1}{2}, 0, 0, 0, 0, \frac{1}{2}\}$ le soddisfa, anche se sappiamo che quella corretta è $\{\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}\}$. Non sapendo quale delle due scegliere, si può sfruttare il **Principio di indifferenza di Laplace** (soddisfatto dalla seconda distribuzione): come calcolare il contenuto di informazione di una distribuzione? Questo principio si applica come regola per assegnare le probabilità a priori. Esso stabilisce che **in assenza di evidenze rilevanti, le probabilità devono essere distribuite equamente tra i gradi di libertà** o esiti possibili. La prima distribuzione ci fornisce più informazione rispetto alla seconda: ci sono solo due esiti possibili, mentre l'altra è "più incerta", con sei esiti possibili equiprobabili. In probabilità Bayesiana questo è il prior più semplice possibile.

Immaginiamo un insieme di N oggetti $\mathcal{M} = \{x_1, \dots, x_N\}$. Facciamo finta che questi siano i biglietti della lotteria: uno solo di questi biglietti è quello vincente. Quante domande mi devo fare per capire qual è il biglietto giusto? In questo caso assumiamo una probabilità uniforme in ciascuno dei biglietti. Chiaramente ci sono tanti modi per scoprirlo: uno poco furbo è fare N domande sul singolo biglietto. "È questo? È quello? È quell'altro? ...". Possiamo fare domande più intelligenti in termini di un algoritmo. Supponiamo per semplicità $N = 2^L$.

- Nel caso in cui la cardinalità $|\mathcal{M}| = 1$ abbiamo finito.
- Altrimenti, costituiamo due insiemi di uguale dimensione $\mathcal{M}_1 = \{x_1, \dots, x_{N/2}\}$ ed $\mathcal{M}_2 = \{x_{N/2+1}, \dots, x_N\}$.
- A questo punto ci chiediamo: l'obiettivo cade nel primo insieme? Se sì allora $\mathcal{M} = \mathcal{M}_1$, altrimenti poniamo $\mathcal{M} = \mathcal{M}_2$, e torniamo al primo punto.

In altre parole, proseguiamo con l'algoritmo spezzettando le domande \rightarrow in questo caso il numero di domande è

$$S = \log_2 N$$

e in questo caso particolare $S = \log_2 2^N = L$, e quindi possiamo pensare a $\log_2 N$ come una misura dell'incertezza nel caso speciale di N eventi equiprobabili. Questo numero (**entropia**) rappresenta il **contenuto di informazione della sequenza**. Per oggetti equiprobabili questo numero ricorda la definizione di Boltzmann dell'entropia

$$S = k_B \ln W,$$

dove W è il numero di microstati disponibili al sistema fisico. Questo mapping è profondo, e può essere mostrato anche nel caso in cui le probabilità non siano equiprobabili.

Consideriamo N biglietti ed m scatole, in cui ogni scatola contiene N/m biglietti. In questo caso possiamo sfruttare una strategia simile: prima calcoliamo la scatola giusta (con $S_1 = \log_2 m$) e poi ci facciamo le domande sul biglietto (con $S_2 = \log_2 (N/m)$). Anche in questo caso

$$S = S_1 + S_2 = \log_2 N.$$

Se invece le scatole hanno un numero diverso di biglietti possiamo dire che

$$\sum_{i=1}^m n_i = N.$$

A questo punto, la probabilità di trovare il biglietto nella scatola i è data da $p_i = n_i/N$, se ci sono più biglietti in una scatola è più probabile vincere scegliendola. Possiamo scrivere l'**entropia di Shannon**

$$S = - \sum_i p_i \log_2 p_i,$$

da cui $S = \log_2 m$. **L'entropia di Shannon rappresenta il contenuto di informazione della distribuzione p_i : l'entropia è massima quando il contenuto di informazione è minimo.** Cambiando punto di vista si può comunque pensare che l'entropia di Shannon dia una misura dell'informazione contenuta in una distribuzione di probabilità. La scelta con meno assunzioni corrisponde al massimo di questa funzione entropia.

Tornando all'esempio del dado, avevamo due tipi di sequenze. L'entropia della prima ci produce

$$S = -2 \times \frac{1}{2} \log_2 \frac{1}{2} + 4 \times 0 = \log_2 2.$$



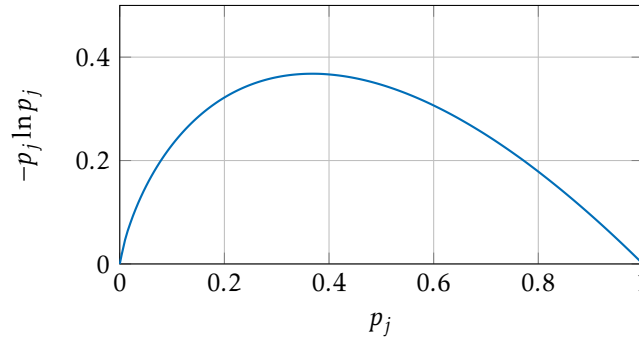


Figura 26: Termine singolo dell'entropia di Shannon.

Nel secondo caso invece

$$S = -6 \times \frac{1}{6} \log_2 \frac{1}{6} = \log_2 6 > \log_2 2.$$

Immaginiamo di avere una certa distribuzione p_i con $i = 1, \dots, N$. Qual è la distribuzione compatibile con MaxEnt consistente con la normalizzazione delle p_i ? Dobbiamo cercare il $\max S(p_i)$ con il vincolo $\sum_i p_i = 1$. Si utilizzano in questo caso i **moltiplicatori di Lagrange**:

$$\max \left(S(p_i) - \lambda \left(\sum_i p_i - 1 \right) \right) \Rightarrow \frac{dS}{dp_i} = 0.$$

Parafrasando l'ultima espressione con la definizione di entropia di Shannon si ottiene

$$\frac{dS}{dp_i} = 0 = -\ln p_i - 1 - \lambda \Rightarrow \ln p_i = -1 - \lambda \Rightarrow p_i = e^{-1-\lambda}.$$

Quindi per la normalizzazione

$$\sum_i e^{-1-\lambda} = N \underbrace{e^{-1-\lambda}}_{p_i} = 1, \Rightarrow p_i = 1/N,$$

distribuzione uniforme, che spesso viene utilizzata come prior. Inoltre, questo valore della probabilità ci fa ottenere un'entropia massima (MaxEnt...) pari a

$$S_{\max} = - \sum_{i=1}^N \frac{1}{N} \ln \left(\frac{1}{N} \right) = \ln N,$$

che corrisponde all'entropia di Boltzmann dell'ensemble microcanonico della meccanica statistica! Infatti MaxEnt è uno dei modi in cui si può dimostrare la PDF di tale ensemble. Questa definizione di entropia ha diverse proprietà:

- $S \geq 0$ perché $0 \leq p_i \leq 1$;
- S è convessa, cioè $\partial^2 S / \partial p_i \partial p_j = -\delta_{ij} / p_i \leq 0$;
- $S = 0$ è deterministico, non c'è incertezza: c'è un'unica configurazione (in accordo e analogia con l'entropia termodinamica);
- in assenza di informazioni testabili, la PDF più probabile è $p_i^{\text{ME}} = \text{const}$, in accordo con il principio di indifferenza di Laplace.

Come si sceglie il prior? Si sfrutta la teoria di E. T. Jaynes, sfruttando MaxEnt e anche due condizioni: normalizzazione $\sum_i p_i = 1$ e soddisfazione di vincoli arbitrari dipendenti dal problema $\Phi_a(\{p\}) = 0$, per esempio media fissata. A questo punto

$$\mathcal{L} = - \sum_i p_i \ln p_i - \lambda_0 \left(\sum_i p_i - 1 \right) - \sum_\alpha \lambda_\alpha \Phi_\alpha(p).$$

Volendo massimizzare,

$$\frac{\partial \mathcal{L}}{\partial p_i} = -\ln p_i - 1 - \lambda_0 - \sum_\alpha \lambda_\alpha \frac{\partial \Phi_\alpha}{\partial p_i} \stackrel{!}{=} 0,$$



da cui si ottiene

$$p_i^{\text{ME}} = \frac{1}{Z} \exp\left(\sum_{\alpha} \lambda_{\alpha} \frac{\partial \Phi_{\alpha}}{\partial p_i}\right), \quad Z = \sum_j \exp\left(\sum_{\alpha} \lambda_{\alpha} \frac{\partial \Phi_{\alpha}}{\partial p_j}\right).$$

Spesso queste equazioni sono irrisolvibili. Un caso risolvibile è quello in cui il vincolo è lineare nelle probabilità

$$\Phi_{\alpha}(p) = \sum_j k_j^{\alpha} p_j,$$

per cui

$$\frac{\partial \Phi_{\alpha}}{\partial p_j} = k_j^{\alpha} \implies p_j^{\text{ME}} = \frac{1}{Z} \exp\left(\sum_{\alpha} k_j^{\alpha} \lambda_{\alpha}\right),$$

con normalizzazione $Z = \sum_i \exp(\sum_{\alpha} k_i^{\alpha} \lambda_{\alpha})$ e vincolo $\sum_j p_j^{\text{ME}} k_j^{\alpha} = 0$.

Un esempio di applicazione al mondo della fisica è quello della distribuzione di Maxwell-Boltzmann: qual è la distribuzione che soddisfa sia $\sum_i p_i = 1$ che il vincolo lineare $\sum_i p_i E_i = \langle E \rangle$ energia media? Data un'energia media nota, qual è la distribuzione che descrive il sistema? Utilizzando MaxEnt troviamo che

$$p_j^{\text{ME}} = \frac{1}{Z} e^{-\beta E_j}, \quad Z = \sum_j e^{-\beta E_j},$$

dove Z è la funzione di partizione e $\langle E \rangle = -\partial \ln Z / \partial \beta$. Questa formula viene dal fatto che

$$\sum_i \frac{1}{Z} e^{-\beta E_i} E_i = E,$$

da cui

$$\frac{\partial \ln Z}{\partial \beta} = \frac{1}{Z} \frac{\partial Z}{\partial \beta} = \frac{1}{Z} \sum_j -E_j e^{-\beta E_j} = -\langle E \rangle.$$

In termodinamica si assume $\beta = 1/k_B T$. **In conclusione, MaxEnt viene sfruttato in probabilità Bayesiana per stabilire l'entità del prior.**

8 Test statistici

Una volta fatta una misura, vogliamo sapere se è compatibile con la teoria, ma dobbiamo capire se sono osservazioni vere o se sono figlie di rumore termico/statistico. Servono metodi statistici forti e convincenti per confermare o meno le osservazioni sperimentale, con conseguente fortissima rilevanza in ambiti come la fisica, la medicina, e in generale in tutti i campi della scienza che hanno a che fare con dei dati.

Il problema è quello di testare un'ipotesi H . In particolare, si cerca sempre (o quasi) di testare un'ipotesi nulla H_0 . Un'ipotesi nulla è un'affermazione sulla distribuzione di probabilità di una o più variabili casuali, e si cerca di assumerla possibilmente vera. Si intende per ipotesi nulla l'affermazione secondo la quale non ci sia differenza oppure non vi sia relazione tra due fenomeni misurati, o associazione tra due gruppi. Solitamente **viene assunta vera, finché non si trova una evidenza che la confuti**. Il fallimento nell'escludere l'ipotesi nulla logicamente NON conferma né supporta tale ipotesi. **Non si dimostra che è vera, ma che non è falsa**. Nel test statistico viene verificata in termini probabilistici la validità di un'ipotesi statistica, detta appunto ipotesi nulla, di solito indicata con H_0 . Attraverso una funzione dei dati campionari si decide se accettare l'ipotesi nulla o meno. Nel caso l'ipotesi nulla venga rifiutata si accetterà l'ipotesi alternativa. Se si rifiuta un'ipotesi nulla che nella realtà è vera, allora si dice che si è commesso un errore di prima specie (o falso positivo). Accettando invece un'ipotesi nulla falsa si commette un errore di seconda specie (o falso negativo). L'ipotesi nulla è che l'effetto che vedo non è l'effetto ma è dovuto al caso, è una fluttuazione.

Possiamo fare un esempio di ipotesi nulla con la solita moneta. Immaginiamo di tirarla N volte, e che otteniamo N volte croce. L'ipotesi nulla H_0 è che la moneta *non* sia truccata—e in questo caso la assumiamo come vera, dobbiamo vedere se si confuta o meno. Qual è la probabilità che l'ipotesi nulla sia valida $p(N|H_0)$? Ovviamente è $p(N|H_0) = (1/2)^N$. Quando decidiamo che la moneta è truccata? Cerchiamo il p -value, cioè probabilità che sia valida H_0 . Più precisamente, esso è probabilità di ottenere risultati uguali o meno probabili di quello osservato durante il test, supposta vera l'ipotesi nulla. In altri termini, il p -value aiuta a capire se la differenza tra il risultato osservato e quello ipotizzato è dovuta alla casualità introdotta dal campionamento, oppure se tale differenza è statisticamente significativa, cioè difficilmente spiegabile mediante la casualità



Importante:

$\Pr(\text{osservazione} \mid \text{ipotesi}) \neq \Pr(\text{ipotesi} \mid \text{osservazione})$

La probabilità di osservare un risultato data per vera una certa ipotesi non è *equivalente* alla probabilità che l'ipotesi sia vera dato un risultato osservato.

Usando il valore-p come “punteggio” si commette un grave errore logico **la fallacia del condizionale trasposto**.

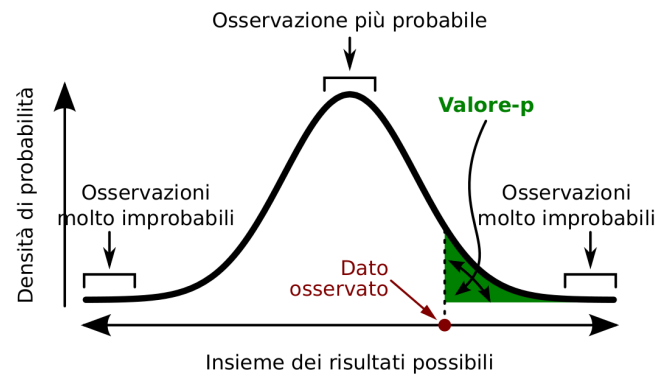


Figura 27: Rappresentazione del p -value (area verde), probabilità del risultato osservato (o più estremo) supponendo sia vera l'ipotesi nulla.

dovuta al campionamento. Questa soglia va decisa *prima* che avvenga l'esperimento, che mi mostra se l'ipotesi nulla è da rigettare o meno. Non mi dirà se l'ipotesi opposta sia vera: ci può dire che è molto improbabile che la moneta (non) sia truccata. Questa soggettività è nota come **livello di significatività** α , tale che se $p < \alpha$ allora H_0 è rifiutata. Nel caso del bosone di Higgs, la particella è stata dichiarata quando il risultato era 5σ dal livello del rumore. Questo valore può essere convertito in p -value tenendo in mente quando detto nella Sec. 6.3. Invece, in biologia si prende solitamente $\alpha = 0.05$. In casi più stringenti si usa 0.01 (“c’è una specie di feticismo per il 0.05”, con conseguente p -value hacking <https://www.youtube.com/watch?v=42QuXLuch3Q>). Nel caso della moneta, per cinque tiri $p = (1/2)^5 = 0.03125$, per cui un biologo direbbe che è truccata (l'ipotesi nulla è stata rigettata, quindi non è vero che non sia truccata), mentre un fisico particellare dovrebbe fare molti più tiri (potrebbe essere un caso cinque esiti uguali consecutivi...).

In generale, quello che si definisce è la cosiddetta **statistica di test**. Immaginiamo di avere un'osservazione $x = \{x_1, \dots, x_N\}$ che vogliamo verificare. Definiamo una funzione $t(x)$ variabile di test. Calcoliamo la PDF che valga $f(t|H_0)$, probabilità su cui facciamo agire la nostra statistica. Ad esempio, possiamo definire $t(x) := \bar{x} = \frac{1}{N} \sum_{i=1}^N x_i$. In questo caso, il p -value sarà la probabilità

$$p = p(t \geq t^* | H_0) = 1 - \int_0^{t^*} dt f(t|H_0).$$

Ci sono casi in cui si fa un'analisi two-sided, con due code $t > t_1^*$ e $t < t_2^*$: si rigetta l'ipotesi nulla se il valore è troppo piccolo o troppo grande rispetto a due valori di riferimento. Ipotesi semplice: $f(t|H_0)$ PDF, mentre le ipotesi complesse hanno pdf . Sia $\alpha = p(x \in R|H_0)$, dove R è la cosiddetta *rejection region* \rightarrow errore di tipo I: rigettiamo un'ipotesi quando questa avrebbe dovuto essere accettata. Invece sia $\beta = p(x \in R|H_1)$ errore di tipo II, probabilità che x appartenga al complementare. L'errore di tipo II avviene quando si accetta un'ipotesi che avrebbe dovuto essere rigettata. Altra nomenclatura importante è la significatività, data da $S = 1 - \alpha$, mentre il potere statistico (a volte anche solo potere) è dato da $PW = 1 - \beta$. Vediamo alcuni esempi di test veri e propri, utilizzati normalmente. È importante ricordare che alcuni di questi test richiedono condizioni a priori (come la gaussianità delle variabili casuali), e che non sempre queste vengono rispettate. L'utilizzo del test adatto alla situazione è fondamentale per qualunque analisi statistica dei dati.



8.1 z-statistics

Supponiamo di avere due set di dati $x_1 = \{x_1^1, \dots, x_1^N\}$ con σ_1^2 come varianza, e $x_2 = \{x_2^1, \dots, x_2^N\}$ con varianza σ_2^2 . L'ipotesi H che vogliamo dimostrare è che $\langle x_1 \rangle = \langle x_2 \rangle$. Per fare ciò definiamo una variabile

$$z := \frac{\bar{x}_2 - \bar{x}_1}{\text{Std-err}^2},$$

dove lo standard error è definito da $\text{Std-err}^2 = \text{SE}^2 = \sigma_1^2/N + \sigma_2^2/N$ dalla composizione delle due varianze. Allora, possiamo rigettare o meno l'ipotesi che la media sia diversa a partire dalla distribuzione di test

$$f(z|H) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2},$$

abbiamo una gaussiana. Allora il p -value è dato dall'integrale

$$1 - \int_{-z}^z dz' f(z'|H).$$

Con un p -value piccolo posso escludere che i campioni abbiano stessa media. Se troviamo un valore maggiore del livello di significatività che ci siamo imposti, allora l'esperimento è inconcludente, e non posso stabilire con l'esperimento se effettivamente è servito a (non) confermare l'ipotesi. Gli altri test sono variazioni sul tema dello Z -test.

8.2 t-test (di Student)

Questo test è uno dei test più utilizzati in biologia, in cui l'ipotesi nulla è H_0 , per cui la media è un valore μ_0 . In questo caso la variabile test è

$$t := \frac{\bar{x} - \mu_0}{\text{SE}},$$

dove lo standard error è $\text{SE} = \sqrt{\Delta x^2/(N-1)}$. Faccio N misure, e voglio vedere se la media è μ_0 . L'altra ipotesi che viene fatta è che x sia una variabile gaussiana. Se le x sono gaussiane, qual è la probabilità che valga questa ipotesi? La soluzione al problema è data dalla funzione t di Student

$$p(t|N, \mu_0) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\pi\nu}} \left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}},$$

dove ν = numero di osservazioni (gradi di libertà)–1, perché uno di questi è stato utilizzato per stimare l'errore standard. È un'altra variante dello z -test. Abbiamo una teoria, facciamo una misura e otteniamo un valore vicino alla teoria. È un valore affidabile o no? Calcolando il p -value dati i dati sperimentali e vediamo se questa probabilità è maggiore o minore del p -value, accettando o meno l'ipotesi che la media sia uguale a μ_0 .

8.3 Test del χ^2

Questo test serve per stabilire la bontà di un fit. Tipicamente si hanno una serie di misure $\vec{y} = (y_1, \dots, y_N)$ e delle variabili di controllo $\vec{s} = (s_1, \dots, s_N)$ —ad esempio in una misurazione della temperatura il funzione della pressione, dove \vec{s} sono le variabili di pressione a cui si misura la temperatura. Qui l'ipotesi è che i dati seguano una certa teoria, per cui H_0 : è l'ipotesi per cui $y_i = f(s_i|\vec{\theta})$, dove abbiamo in principio dipendenza da più parametri $\vec{\theta} = (\theta_1, \dots, \theta_R)$. La statistica del χ^2 è definita dalla variabile

$$\chi := \sum_{i=1}^N \frac{(y_i - f(s_i|\vec{\theta}))^2}{\sigma_i^2},$$

per cui

$$f(\chi|H_0, \nu) = \frac{2^{-\frac{\nu}{2}} \chi^{\frac{\nu}{2}-1} e^{-\frac{\chi}{2}}}{\Gamma(\nu)}, \quad \nu = N - R.$$

Anche qui ci sono delle ipotesi: per costruire questa funzione f si fa l'assunzione che questa misura sia uguale alla teoria a meno di fluttuazioni gaussiane. Di fatto abbiamo assunto che gli errori di misura sono gaussiani, cosa che non necessariamente è vera.



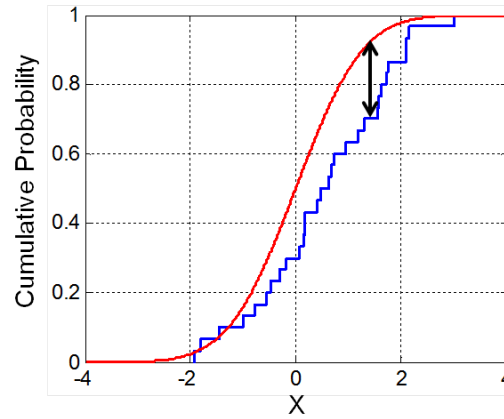


Figura 28: Illustrazione del test di Kolmogorov-Smirnov. La linea rossa rappresenta una cumulata teorica, quella blu una sperimentale, e la freccia nera è la statistica KS.

8.4 Test di Kolmogorov-Smirnov

Questo test è estremamente importante in statistica poiché è **privo di assunzioni sulle variabili**. Tutti questi test fatti prima assumono la gaussianità delle fluttuazioni, che è comunque un'assunzione forte (ad esempio il rumore Barkhausen è descritto da una statistica Gamma, cfr. Sec. 6.4). In questo si mostra la forza del test di Kolmogorov-Smirnov. Assumiamo quindi $\vec{x} = \{x_1, \dots, x_N\}$ e H ipotesi che la PDF sia $f_0(x)$. Quindi la distribuzione cumulata è

$$F_0(x) = \int_{-\infty}^x dx' f_0(x'),$$

e l'idea è costruire una cumulata sperimentale da confrontare con quella teorica. Si calcolano due probabilità D_- e D_+ , distanze rispettivamente minima e massima tra le due curve, e poi si calcola

$$D = \max(D_+, D_-).$$

Successivamente, con un calcolo di tipo RW si calcola la distribuzione per queste D , ottenendo la distribuzione di Kolmogorov per N osservazioni

$$p_K(x \geq D_N \sqrt{N}) \rightarrow \frac{\sqrt{2\pi}}{x} \sum_{n=0}^{\infty} e^{-(2n-1)^2 \pi^2 / 8x^2}.$$

Questo test è comodo quando per esempio si hanno due campioni di distribuzione non specificata. Supponiamo $\{x_1, \dots, x_N\}$ e $\{y_1, \dots, y_N\}$. Sono distribuiti dalla stessa distribuzione o no? Definiamo $D^* = D_N \sqrt{N}$, con $N^* = 1/(\frac{1}{N_1} + \frac{1}{N_2})$. Le applicazioni sono svariate, in primis vedere la dipendenza da un parametro o meno in un certo esperimento (temperatura, pressione, ...).

9 Principal component analysis

Questa è un metodo standard di riduzione dimensionale proveniente dal mondo della data science/statistica. Spesso, nello studio statistico dei dati abbiamo a che fare con tantissime variabili e altrettante caratteristiche: per una certa misura di un materiale possiamo misurare la temperatura di fusione, la resistenza elettrica, la suscettività magnetica e così via. Si possono fare queste misure per tanti materiali. In generale abbiamo una matrice y_{np} dove $n = 1, \dots, N$ i valori (come il tipo di materiale) e $p = 1, \dots, P$ le caratteristiche. Quando dobbiamo analizzare un oggetto di questo tipo, se P è grande, è difficile realizzare quali siano quelle più importanti e se ci siano eventuali pattern. Sono tutte indipendenti queste caratteristiche? Magari non c'è bisogno di guardarle tutte, e possiamo ridurre il numero di incognite studiando le combinazioni di caratteristiche che descrivono il sistema. Se sono correlate, basta trovare il fattore principale (Fig. 29). Riassumendo, scopo della tecnica è quello di **ridurre il numero più o meno elevato di variabili che descrivono un insieme di dati a un numero minore di variabili latenti, limitando il più possibile la perdita di informazione**.

Solitamente come prima cosa si normalizzano le variabili, secondo

$$x_{np} := \frac{y_{np} - \bar{y}_p}{\delta_p},$$



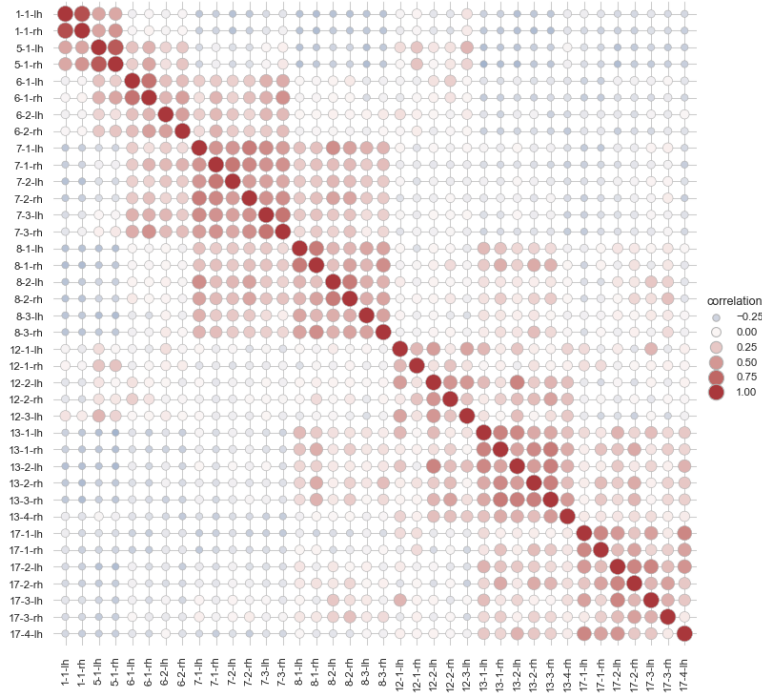


Figura 29: Esempio di matrice di correlazione sovrapposta ad uno scatterplot. Si osservi come la diagonale sia auto-correlata in maniera banale.

dove

$$\bar{y}_p = \frac{1}{N} \sum_n y_{np}, \quad \text{e} \quad \delta_p^2 = \frac{1}{N-1} \sum_{n=1}^N (y_{np} - \bar{y}_p)^2.$$

In questo modo

$$\langle x_{np} \rangle = 0, \quad \text{e} \quad \langle (\Delta x_{np})^2 \rangle = 1.$$

Scriviamo la matrice di covarianza $C = \frac{1}{N-1} x^T x$, tale che

$$C_{pq} = \sum_{n=1}^N \frac{x_{np} x_{nq}}{N-1}.$$

La matrice è simmetrica, per cui $C_{pq} = C_{qp}$, ed essendo simmetrica può essere diagonalizzabile. Possiamo trasformarla tramite una matrice V tale che $C \mapsto V^T C V = D$, con

$$D = \text{diag}(\lambda_1, \dots, \lambda_p),$$

dove λ_i sono gli autovalori di C . Questa diagonalizzazione si ottiene nel solito modo

$$(C - \lambda_p I) v_p = 0 \iff C v_p = \lambda_p v_p,$$

oppure annullando il $\det(C - \lambda_p I) = 0$. Conoscendo la matrice V che diagonalizza quella delle correlazioni, possiamo definire delle variabili $z_{np} = V^T x_{np}$, tali che $x_{np} = V z_{np}$.

Nella PCA si guardano i fattori $f_{np} := z_{np} / \sqrt{\lambda_p}$, tale che

$$\langle f_{np} \rangle = 0 \quad \text{e} \quad \langle (\Delta f_{np})^2 \rangle = 1.$$

Gli autovalori maggiori avranno peso maggiore, e saranno più importanti. Questo è collegato alla analisi fattoriale, per cui se uno ha delle variabili x_1, \dots, x_p che vuole scrivere come combinazione lineare di una serie di fattori come

$$x_1 = a_{11} f_1 + \dots + a_{1q} f_q + \varepsilon_1$$

$$\vdots$$

$$x_p = a_{p1} f_1 + \dots + a_{pq} f_q + \varepsilon_p$$



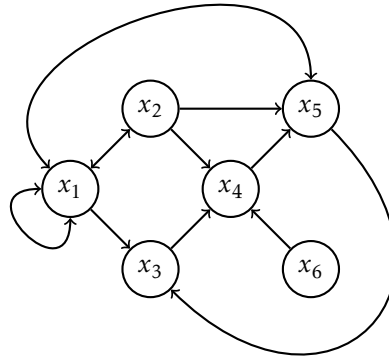


Figura 30: Rappresentazione di un network diretto a sei nodi $\{x_1, \dots, x_6\}$.

In questo caso $q < p$, e ε_i rappresenta l'errore su questa approssimazione. **Gli autovalori più grandi di queste combinazioni lineari rappresentano meglio il sistema.** In questo modo posso capire quali siano le variabili che pesano di più nell'analisi dei dati.

Un'altra tecnica che si usa per trovare questa decomposizione dei fattori è la cosiddetta *singular value decomposition* (SVD). L'idea questa volta è di trovare la matrice (che non è quadrata, e quindi non diagonalizzabile)

$$x_{np} = UDV^T.$$

Anche in questo modo si possono trovare i fattori e quindi gli pseudo-autovalori di questa matrice. È importante sapere che esistono degli algoritmi matematici che fanno questa procedura in automatico per qualunque matrice.

10 Networks

Spesso si sente parlare del fatto che molte relazioni possono essere rappresentate mostrandole come una serie di connessioni—esempi sono i social networks, oppure l'internet. Network è sinonimo di grafo, nel senso matematico della parola. Sono la stessa cosa, ma in due contesti diversi. Vogliamo mostrare come utilizzare i networks come mezzo di rappresentazione delle correlazioni tra variabili.

In generale un network possiamo immaginarlo come una serie di punti connessi da legami (link). Abbiamo una serie di nodi $i = 1, \dots, N$ e una serie di edges $l = 1, \dots, L$. Un network può essere rappresentato da una matrice di adiacenza, che ci dice “chi è vicino a chi” A_{ij} , di entrate 1 e 0. I grafi non diretti non hanno frecce ma solo linee, per cui $A_{ij} = A_{ji}$. I grafi diretti (Fig. 30) hanno invece $A_{ij} \neq A_{ji}$, e le frecce possono essere a doppia direzione. Dato un grafo, si possono definire una serie di quantità, come ad esempio il grado (degree), variabile definita per ogni nodo

$$K_i = \sum_{j=1}^N A_{ji}$$

per il grafo non diretto. Nel caso diretto bisogna differenziare due tipi di gradi:

$$K_i^{(\text{in})} = \sum_{j=1}^N A_{ij}, \quad K_i^{(\text{out})} = \sum_{j=1}^N A_{ji}.$$

Il grado totale è dato da

$$K_i^{\text{tot}} = K_i^{(\text{in})} + K_i^{(\text{out})}.$$

Vale la proprietà che

$$\sum_i K_i^{(\text{in})} = \sum_i K_i^{(\text{out})} = 2L = \sum_{ij} A_{ij}.$$



Dato un network si possono studiare le sue proprietà statistiche, come il suo grado medio. Questo è dato da

$$\langle K \rangle = \frac{1}{N} \sum_i K_i = \frac{2L}{N}.$$

Altra cosa che si può studiare è la distribuzione dei gradi: in un reticolo quadrato il grado è 4, e la distribuzione è una delta centrata su 4. In generale non è così, e alcuni grafi hanno una distribuzione con code molto larghe (distribuzioni a potenza, tipiche dei power networks). Queste hanno proprietà non banali: ci sono pochi siti con tantissime connessioni, e la maggior parte dei siti ne ha pochi, per cui

$$p_k \sim 1/K^\alpha.$$

Esistono anche dei network pesati: in generale può essere associato un peso ad ogni link, associando una linea spessa a piacere a seconda del valore della connessione. Se pensiamo ad internet, ci si può chiedere quanto traffico passi tra i nodi i e j .



Elenco delle figure

1	Funzione della probabilità di massa della somma di due dadi.	2
2	Rappresentazione geometrica del Teorema di Bayes sul quadrato unitario 1×1 . Osserviamo come $P(E H)$ sia dato dalla <i>proporzione</i> di $P(E)$ rispetto a $P(H)$	3
3	Esempio più grande del paradosso di Monty Hall con 20 porte, 19 capre e 1 macchina.	4
4	Cumulativa di diverse distribuzioni gaussiane ed esempio di una funzione di sopravvivenza. . .	5
5	PDF di tre diverse distribuzioni (normalizzate): lifetime distribution con $\tau = 1$, uniforme, e gaussiana di media nulla $x_0 = 0$ e deviazione standard unitaria $s = 1$	7
6	Esempio di una mappa non lineare $f : \mathbb{R}^2 \mapsto \mathbb{R}^2$ manda un piccolo quadrato in un parallelogramma distorto. Lo jacobiano dà la migliore approssimazione lineare del parallelogramma distorto, e il suo determinante dà il rapporto tra l'area del parallelogramma e quella del quadrato originario.	10
7	PDF della distribuzione binomiale. Si osservi come si allarghi la distribuzione all'aumentare del numero di campioni n (a parità di probabilità).	12
8	PDF e cumulativa di più distribuzioni poissoniane a confronto.	15
9	Distribuzione di Poisson come limite della distribuzione binomiale al variare del numero di tiri. . .	16
10	PDF e cumulativa di più distribuzioni gaussiane a confronto.	17
11	Distribuzione del χ^2 e rispettive cumulative.	18
12	Alcune distribuzioni Gamma e rispettive cumulative, dove $\nu = k$ e $\lambda = 1/\theta$	19
13	Distribuzioni Beta e rispettive cumulative.	20
14	Distribuzioni di Cauchy e Levy a confronto.	21
15	Distribuzioni geometriche e rispettive cumulative.	23
16	Distribuzioni ipergeometriche e relative cumulative.	23
17	Rappresentazione del teorema del limite centrale. A sinistra, il classico esempio in cui cadono una serie di biglie con pari probabilità di andare a destra o sinistra ad ogni bivio: la distribuzione risultante quando vengono collezionate viene approssimata sempre più precisamente da una gaussiana mano a mano che le biglie cadono.	25
18	Rappresentazione di una bigaussiana di variabili scorrelate.	29
19	PDF e la corrispondente curva di livello di due variabili gaussiane correlate con $\rho = 1/2$, $s_x = 1$ e $s_y = 1.4$. Tale correlazione tra le variabili ruota l'asse dell'ellisse (in grigio) di un angolo $\varphi = \arctan(1.46) \approx 50^\circ$	30
20	Sopra, PDF e rispettive cumulative della distribuzione di Gumbel, dove $\mu = \alpha$. Sotto, simulazioni della distribuzione di $\langle M \rangle$ presi i massimi da distribuzioni gaussiane di media e varianza variabili, confrontate con il valore teorico $\langle M \rangle = \frac{1}{2} \ln N$	32
21	PDF e rispettive cumulative della distribuzione di Weibull.	33
22	Esempio di un boxplot, con rappresentati alcuni outliers come \bullet	34
23	Confronto tra likelihoods. A sinistra, la likelihood per la probabilità p_T^2 che una moneta tirata due volte faccia testa-testa (senza conoscenza a priori sul fatto che sia truccata o meno). A destra, la likelihood per la probabilità $p_T^2(1 - p_T)$ per un'osservazione testa-testa-croce, sempre senza prior.	35
24	Rappresentazione degli intervalli di confidenza di una distribuzione gaussiana.	39
25	Ciclo di isteresi e rumore di Barkhausen.	40
26	Termine singolo dell'entropia di Shannon.	43
27	Rappresentazione del p -value (area verde), probabilità del risultato osservato (o più estremo) supponendo sia vera l'ipotesi nulla.	45
28	Illustrazione del test di Kolmogorov-Smirnov. La linea rossa rappresenta una cumulata teorica, quella blu una sperimentale, e la freccia nera è la statistica KS.	47
29	Esempio di matrice di correlazione sovrapposta ad uno scatterplot. Si osservi come la diagonale sia auto-correlata in maniera banale.	48
30	Rappresentazione di un network diretto a sei nodi $\{x_1, \dots, x_6\}$	49