Distribuzioni “famose” di variabili aleatorie e applicazioni biotecnologiche

Premessa

Il seguente elaborato utilizza le nozioni sulle probabilità fin qui apprese. Lo scopo è quello di spiegare, o almeno tentare, nuove distribuzioni di variabili aleatorie, oltre alle già conosciute bernoulliana e binomiale, ed evidenziare come queste si ritrovino nella pratica quotidiana del biotecnologo. Il Paper seguente è volutamente più tecnico della presentazione e con richiami ad argomenti già fatti, sperando, qualora anche nel corso del prossimo anno si volesse affrontare l'argomento, di rendere il documento anche fonte da cui attingere formule utili.

Richiami alla media e alla varianza in statistica descrittiva

La **media** è un indice di posizione. Come si calcola la media di una distribuzione?

$$\overline{X} = \frac{1}{N}\sum_{i=1}^{n} x_i f_i = \sum_{i=1}^{n} x_i p_i.$$

N rappresenta la numerosità della popolazione, *n* il numero di classi, *x* il valore assegnato alla classe, *f* la frequenza assoluta per ogni classe e *p* la frequenza relativa. Per chi non ha simpatia per il simbolo di sommatoria possiamo riscrivere, per esteso:

$$\overline{X} = \frac{1}{N}(x_1\, f_1 + x_2\, f_2 + \ldots + x_n\, f_n) = (x_1\, p_1 + x_2\, p_2 + \ldots + x_n\, p_n).$$

Risulta anche interessante misurare il grado di dispersione dei dati rispetto alla media. Si osservi che la somma delle deviazioni dalla media è sempre zero, ovvero

$$\sum_{i=1}^{N}(x_i - \overline{X}) = 0,$$

perciò, per misurare in modo significativo la dispersione dei dati rispetto alla media, si può considerare la somma dei moduli delle deviazioni, oppure la somma dei quadrati delle deviazioni. Si chiama **varianza** di X e si indica con s^2_X, la media degli scarti al quadrato, ovvero la quantità

$$s^2_X = \frac{1}{N}\sum_{i=1}^{n}(x_i - \overline{X})^2 f_i = \sum_{i=1}^{n}(x_i - \overline{X})^2 p_i.$$

Per il calcolo esplicito della varianza è utile la seguente formula

$$s^2_X = \frac{1}{N}\sum_{i=1}^{N} x_i^2\, f_i - \overline{X}^2 = \overline{X^2} - \overline{X}^2.$$

Si chiama **scarto quadratico medio** o anche **deviazione standard** di X e si indica con s_X, la radice della varianza, ovvero

$$s_X = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{X})^2 f_i} = \sqrt{\sum_{i=1}^{N}(x_i - \overline{X})^2 p_i}.$$

Media e varianza di una variabile aleatoria

In analogia a quanto fatto per le distribuzioni di frequenze possiamo introdurre la media e la varianza di una variabile aleatoria. Si chiama **media** di X e si indica con μ_X o E[X] (la E sta per expectation, in inglese aspettazione o media), la quantità

$$\mu_x = \mathbb{E}[X] = \sum_{i=1}^{n} x_i p_i = x_1 \cdot p_1 + x_2 \cdot p_2 + \ldots x_n \cdot p_n.$$

Questa media è valida per variabili aleatorie finite. Una variabile aleatoria X si dice finita se l'insieme E è un insieme finito. E è l'insieme dei valori assunti da X (l'immagine di X). Quindi si ha $P(X = x_i) = p_i$, $i = 1, 2, \ldots . n$. Riportando in una tabella si ha

X	x_1	x_2	$\ldots$	x_n
Prob	p_1	p_2	$\ldots$	p_n

Si chiama **varianza** di X e si indica con σ^2_X o Var[X], la quantità

$$\mathrm{Var}[X] = \sum_{i=1}^{n} (x_i - \mu_X)^2 p_i = (x_1 - \mu_X)^2 \cdot p_1 + (x_2 - \mu_X)^2 \cdot p_2 + \ldots (x_n - \mu_X)^2 \cdot p_n.$$

Si può anche scrivere così

$$\sigma_X^2 = \sum_{i=1}^{n} x_i^2 p_i - \mu_X^2 = \mathbb{E}[X^2] - \mathbb{E}[X]^2.$$

E la **deviazione standard** di conseguenza è

$$\sigma_X = \sqrt{\sum_{i=1}^{n} (x_i - \mu_X)^2 p_i}.$$

Tuttavia esistono altre formalizzazioni della media e della varianza quando E è un insieme numerabile ma non finito oppure quando E è continuo.

La distribuzione di Bernoulli

La variabile aleatoria con distribuzione di *Bernoulli* è, forse, la più semplice, ma anche una delle più importanti. Consideriamo un esperimento con due soli esiti; definiamo successo uno dei due esiti, insuccesso l'altro. Sia p la probabilità di successo, quindi 1 - p è la probabilità di insuccesso. A questo punto sia X la variabile aleatoria che vale 1 se ottengo il successo, 0 se ottengo l'insuccesso. Si ha

$$\mathbb{P}(X = 1) = p \qquad \mathbb{P}(X = 0) = 1 - p.$$

Sotto forma di tabella

X	0	1
Prob	$1 - p$	p

Calcoliamo la media e la varianza

$$\mathbb{E}[X] = 1 \cdot p + 0 \cdot (1 - p),$$

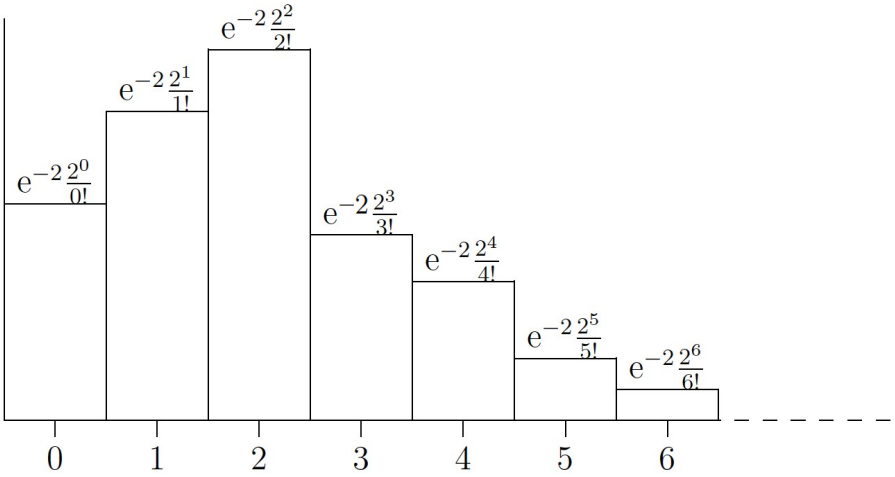
$$\mathrm{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = p - p^2 = p(1 - p).$$

La distribuzione di Poisson

La variabile aleatoria X ha legge di *Poisson* di parametro $\lambda > 0$, se X ha la seguente distribuzione

$$\mathbb{P}(X = k) = \mathrm{e}^{-\lambda}\frac{\lambda^{\mathrm{k}}}{\mathrm{k!}}, \quad \text{per} \quad \mathrm{k} = 0, 1, \ldots .$$

Questa distribuzione infinitamente numerabile si manifesta in molti fenomeni naturali, come: il numero di chiamate ad un centralino in un'unità di tempo (minuto, ora, giorno, ecc), il numero di auto che transitano in un certo incrocio sempre in un'unità di tempo e altri fenomeni simili. Segue un esempio di diagramma della distribuzione di Poisson.



Ci sono 2 errori nella distribuzione. Chi li trova?
La media e la varianza:

$$\mathbb{E}[X] = \lambda, \qquad \mathrm{Var}[X] = \lambda.$$

Ovviamente sono presenti molte altre distribuzioni che rivestono un ruolo di primo piano nel calcolo delle probabilità, tra cui: la binomiale (già studiata),la geometrica, la esponenziale.

La distribuzione Normale o Gaussiana

La distribuzione Normale o Gaussiana è una delle più importanti distribuzioni utilizzate in statistica per diversi motivi. Essa è una distribuzione continua con valori su tutto R. Molte variabili casuali nella realtà (la quantità di pioggia che cade in una certa regione, altezza di persone coetanee nella stessa popolazione, peso di persone coetanee nella stessa popolazione, ecc...) seguono una distribuzione Normale. Inoltre la distribuzione Normale serve per approssimare molte altre distribuzioni, come si vedrà successivamente. Diremo che X è una variabile aleatoria Normale di parametri $\mu \in \mathrm{R}$ e $\sigma^2 > 0$, scrivendo $\mathrm{X} \sim \mathrm{N}(\mu;\sigma^2)$, ha una funzione

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad x \in \mathbb{R}.$$

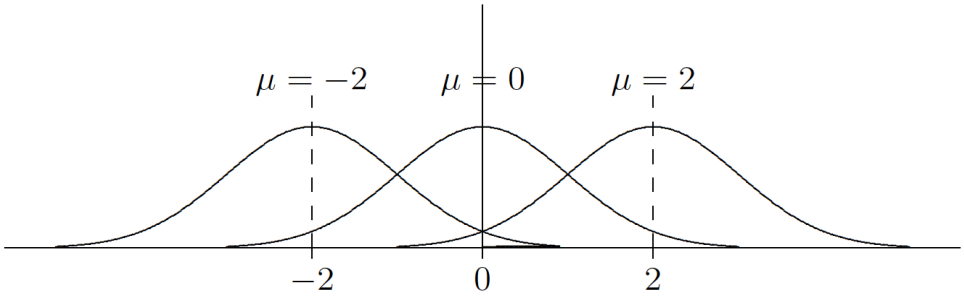
Con $\mathrm{X} \sim \mathrm{N}(\mu;\sigma^2)$

$$\mathbb{E}[X] = \mu, \qquad \mathrm{Var}[X] = \sigma^2.$$

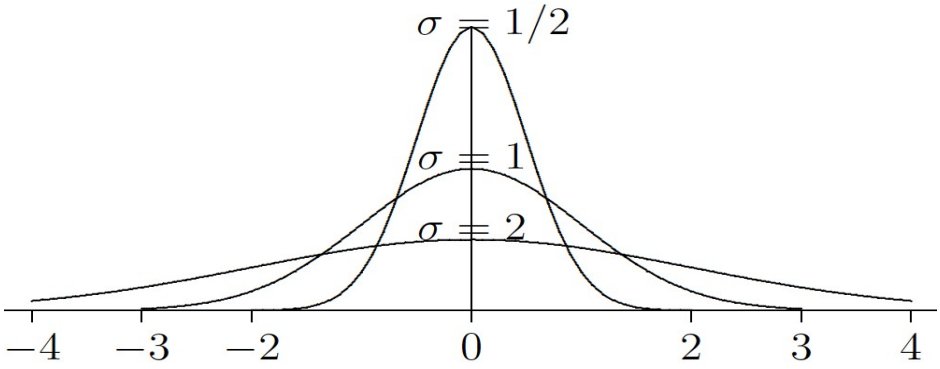
Si noti che queste curve campaniformi sono simmetriche rispetto alla retta $\mu = x$.

Nelle prossime due figure si renderà chiaro su cosa influiscano μ e σ^2.

Distribuzioni normali con σ = 1 fisso, al variare di μ.



Distribuzioni normali con μ = 0 fisso, al variare di σ.



Si può osservare dalle figure sopra, che al variare di μ la campana viene soltanto traslata mantenendo la stessa forma, sempre simmetrica rispetto alla retta $x = \mu$. Al variare di σ, invece la campana si modifica, precisamente è più *concentrata* vicino alla media per valori piccoli di σ e molto più *dispersa* per valori grandi di σ, il che è ovvio se si pensa al significato del parametro σ.
Nel caso delle distribuzioni continue non è utile pensare a quale sarà il la probabilità per un singolo valore, ora ad ogni singolo valore che viene preso in considerazione non corrisponde più l'area di un rettangolo, ma un segmento di cui l'area risultante è 0. Piuttosto bisogna pensare per intervalli e quindi a quale sarà la probabilità che $m < X < n$, con m e n valori qualsiasi dell'asse x. Supponiamo di voler calcolare $P(1 < X < 2)$. Per quanto ne sappiamo sino ad ora questo è pari all'area sottesa dalla funzione (o densità) gaussiana tra $x = 1$ e $x = 2$, quindi è

$$\mathbb{P}(1 < X < 2) = \int_1^2 \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-x^2/2}\, dx.$$

Aiuto! Come si risolve quest'integrale definito? Per fortuna c'è chi ha escogitato un modo per evitare la risoluzione dell'integrale. Infatti per $X \sim N(0;1)$, cioè la *Normale Standard*, esiste una tabella da cui ricavare il risultato. Ma se la distribuzione che si sta studiando non è $X \sim N(0;1)$? Esiste un processo che permette di passare da una Gaussiana qualunque ad una gaussiana standard che si chiama, appunto, *standardizzazione*. Con $X \sim N(0;1)$ è valida

$$X^* = \frac{X - \mu}{\sigma} \sim \mathrm{N}(0, 1).$$

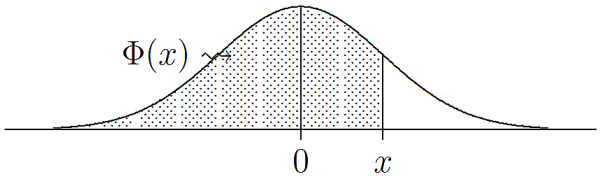
Per fare un esempio: Supponiamo che la temperatura T nel mese di giugno sia distribuita normalmente con media $\mu = 20^oC$ e scarto quadratico medio $\sigma = 5^oC$. Vogliamo determinare la probabilità che la temperatura sia minore di 15^oC; Prendendo in considerazione la *standardizzazione* si ha

$$T^* = \tfrac{T-20}{5} \sim \mathrm{N}(0, 1)$$

$$\mathbb{P}(T < 15) = \mathbb{P}\left(\frac{T-20}{5} < \frac{15-20}{5}\right) = \mathbb{P}(T^* < -1)$$

Ora si guarda la tabella:

Area sottesa dalla Gaussiana Standard ($\Phi(x)$ è l'area ombreggiata)

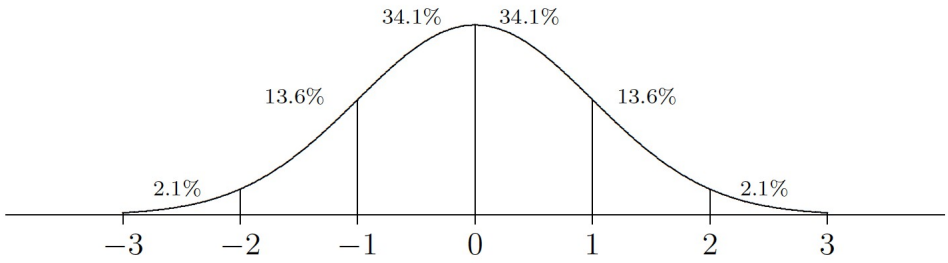


x	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.50000	.50399	.50798	.51197	.51595	.51994	.52392	.52790	.53188	.53586
0.1	.53983	.54380	.54776	.55172	.55567	.55962	.56356	.56750	.57142	.57535
0.2	.57926	.58317	.58706	.59095	.59483	.59871	.60257	.60642	.61026	.61409
0.3	.61791	.62172	.62552	.62930	.63307	.63683	.64058	.64431	.64803	.65173
0.4	.65542	.65910	.66276	.66640	.67003	.67364	.67724	.68082	.68439	.68793
0.5	.69146	.69497	.69847	.70194	.70540	.70884	.71226	.71566	.71904	.72240
0.6	.72575	.72907	.73237	.73565	.73891	.74215	.74537	.74857	.75175	.75490
0.7	.75804	.76115	.76424	.76731	.77035	.77337	.77637	.77935	.78230	.78524
0.8	.78814	.79103	.79389	.79673	.79955	.80234	.80511	.80785	.81057	.81327
0.9	.81594	.81859	.82121	.82381	.82639	.82894	.83147	.83398	.83646	.83891
1.0	.84134	.84375	.84614	.84850	.85083	.85314	.85543	.85769	.85993	.86214
1.1	.86433	.86650	.86864	.87076	.87286	.87493	.87698	.87900	.88100	.88298
1.2	.88493	.88686	.88877	.89065	.89251	.89435	.89617	.89796	.89973	.90147
1.3	.90320	.90490	.90658	.90824	.90988	.91149	.91309	.91466	.91621	.91774
1.4	.91924	.92073	.92220	.92364	.92507	.92647	.92786	.92922	.93056	.93189
1.5	.93319	.93448	.93574	.93699	.93822	.93943	.94062	.94179	.94295	.94408
1.6	.94520	.94630	.94738	.94845	.94950	.95053	.95154	.95254	.95352	.95449
1.7	.95543	.95637	.95728	.95819	.95907	.95994	.96080	.96160	.96246	.96327
1.8	.96407	.96485	.96562	.96638	.96712	.96784	.96856	.96926	.96995	.97062
1.9	.97128	.97193	.97257	.97320	.97381	.97441	.97500	.97558	.97615	.97670
2.0	.97725	.97778	.97831	.97882	.97933	.97982	.98030	.98077	.98124	.98169
2.1	.98214	.98257	.98300	.98341	.98382	.98422	.98461	.98500	.98537	.98574
2.2	.98610	.98645	.98679	.98713	.98745	.98778	.98809	.98840	.98870	.98899
2.3	.98928	.98956	.98983	.99010	.99036	.99061	.99086	.99111	.99134	.99158
2.4	.99180	.99202	.99224	.99245	.99266	.99286	.99305	.99324	.99343	.99361
2.5	.99379	.99396	.99413	.99430	.99446	.99461	.99477	.99492	.99506	.99520
2.6	.99534	.99547	.99560	.99573	.99585	.99598	.99609	.99621	.99632	.99643
2.7	.99653	.99664	.99674	.99683	.99693	.99702	.99711	.99720	.99728	.99736
2.8	.99745	.99752	.99760	.99767	.99774	.99781	.99788	.99795	.99801	.99807
2.9	.99813	.99819	.99825	.99831	.99836	.99841	.99846	.99851	.99856	.99861
3.0	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

La tabella fornisce l'area sottesa dalla curva normale standard minore di un x positivo. La simmetria della curva rispetto all'asse ad $x = 0$ consente di ricavare l'area compresa tra due valori qualunque di x. Se indichiamo con $\Phi(X)$ il valore dell'area sottesa sino ad x (ovvero il valore fornito dalla tabella), possiamo allora rispondere alla questione prima posta, ovvero quanto fa P(T* < -1)?

$$\mathbb{P}(T^* < -1) = 1.000 - \Phi(1) = 1.000 - 0.8413$$

Esaminando un po' più da vicino il grafico della Normale Standard, possiamo vedere che per -1 < x < 1 si ha il 68.2% della distribuzione e per -2 < x < 2 ben il 95.4%, e per -3 < x < 3 praticamente il 100%.

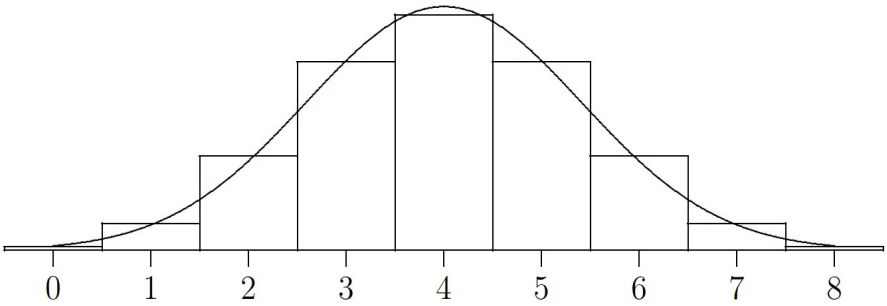


Il Teorema Limite Centrale

É importante menzionare il risultato di un teorema, importante nel calcolo delle probabilità e per le sue applicazioni in statistica: il *Teorema Limite Centrale.*
In termini semplicistici esso afferma che la somma di un *gran* numero di variabili aleatorie tutte con la stessa distribuzione *tende* ad avere una distribuzione normale. L'importanza di ciò sta nel fatto che siamo in grado di ottenere stime della probabilità che riguardano la somma di variabili aleatorie indipendenti ed identicamente distribuite (i.i.d), indipendentemente da quale sia la distribuzione di ciascuna. Ripensando quindi alla binomiale come somma di variabili aleatorie ($X = Y_1 + Y_2 + \ldots + Y_n$) e con ogni variabile aleatoria (Y) avente una distribuzione bernoulliana, si può pensare di applicare il Teorema Limite Centrale per valori (n) grandi. Quindi la media e la varianza della distribuzione binomiale possono essere pensate come

$$\mathbb{E}[X] = \mathbb{E}[Y_1 + Y_2 + \ldots Y_n] = \mathbb{E}[Y_1] + \mathbb{E}[Y_2] + \ldots + \mathbb{E}[Y_n] = p + p + \ldots + p = np,$$

$$\begin{aligned}\mathrm{Var}[X] &= \mathrm{Var}[Y_1 + Y_2 + \ldots Y_n] = \mathrm{Var}[Y_1] + \mathrm{Var}[Y_2] + \ldots + \mathrm{Var}[Y_n] \\ &= p(1-p) + p(1-p) + \ldots + p(1-p) = np(1-p).\end{aligned}$$



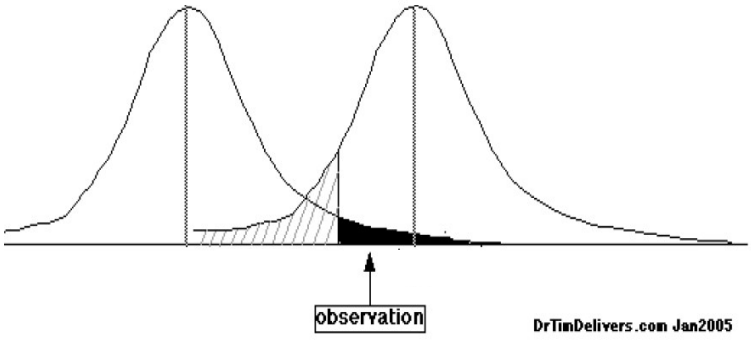
É buona norma quella di applicare l'approssimazione normale della binomiale solo se

$$np > 5, \quad n(1-p) > 5.$$

La distribuzione gaussiana nella pratica biologica e medica

La distribuzione gaussiana descrive molte situazioni reali, come precedentemente accennato. Ma, oltre a esempi banali, se ne possono fare altri di maggiore interesse. Uno di questi riguarda le analisi di laboratorio di campioni organici (sangue, urine, ecc...). Dopo aver raccolto un numero elevato di campioni in una popolazione di individui sani, si nota che i valori statistici, ad esempio di glicemia, vanno a costituire la distribuzione già vista nella binomiale, ovviamente non è possibile avere subito una distribuzione continua perché i metodi analitici hanno un limite di sensibilità. All'aumentare dei

campioni si vede come la media tenda a stabilizzarsi intorno a un valore che risulterà, per il fatto che si può pensare alla binomiale (per un alto n) come una distribuzione gaussiana, pari alla media (μ). Per cui un individuo, di cui si sa che è sano, avrà una certa probabilità di avere un determinato valore in accordo con la gaussiana.
I risultati delle analisi sono spesso accompagnati da valori di riferimento che determinano un intervallo in cui la persona esaminata presenta una situazione di normalità. Tuttavia non è impossibile che una persona sana abbia valori al di fuori di quelli di riferimento: è solo poco probabile. Il genoma e le condizioni ambientali concorrono per la spiegazione.
Quindi, nel caso in cui ci si trovasse a fare degli studi su cavie degli effetti di un farmaco sperimentale, è indispensabile fare degli studi preliminari per evitare poi di ritrovarsi con dei risultati fuorvianti che indicano una azione del composto non veritiera.
Tornando al discorso sui valori di riferimento; anche la popolazione di individui malati può seguire una distribuzione gaussiana. Ovviamente se le aree descritte dalle curve, dei sani e dei malati, sono coincidenti, il test non è considerato indicativo o determinante per la diagnosi. Talvolta le distribuzioni si sovrappongono in piccola parte, in tal caso viene effettuato un taglio (cut off) all'interno dell'intervallo di valori coperto da entrambe le distribuzioni; da una parte la popolazione è sana, dall'altra è malata. Così si vengono a crearsi dei falsi positivi o falsi negativi ai test. La scelta su dove effettuare il taglio determina la sensibilità o la specificità dell'esame.

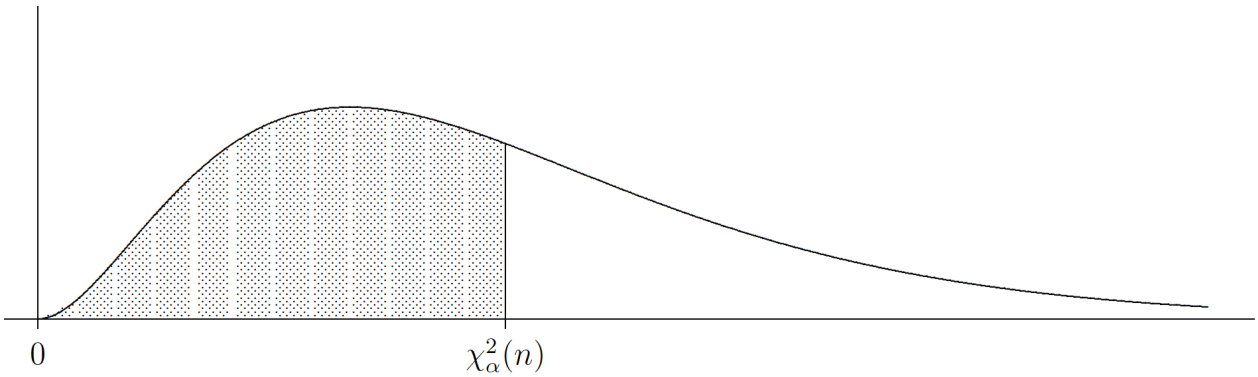


La distribuzione χ^2 (chi quadro)

La variabile aleatoria X ha distribuzione χ^2 con n gradi di libertà e scriveremo $X \sim \chi^2(n)$ se X ha densità data da

$$f(x) = c_n x^{n/2-1} \mathrm{e}^{-x/2}, \quad \text{per} \quad x > 0,$$

dove c_n è una costante opportuna.
Questo è l'andamento tipico di una distribuzione di questo tipo (qui n = 15)



Anche qui, come per la distribuzione gaussiana, si usa una tabella opportuna (qui non riportata).
Si chiama $\chi^2(n)$ quel valore sull'asse *x* tale che:

$$\mathbb{P}(X < \chi^2_\alpha(n)) = \alpha, \quad se \quad X \sim \chi^2(n), \quad \alpha \in (0,1).$$

Per n maggiore o uguale a 30 esiste una formula per convertire i quantili della distribuzione del χ^2 nei quantili della distribuzione di Gauss.

$$\chi^2_\alpha(n) \simeq z_\alpha \sqrt{2n} + n$$

Il test del χ^2 di adattamento

É una importante procedura statistica che ha lo scopo di verificare se certi dati empirici si adattino bene ad una distribuzione teorica assegnata. Permette di confutare l'esistenza di una associazione tra dati e distribuzione teorica, qualora la probabilità che ciò sia vero è troppo bassa per essere accettata come una fluttuazione casuale dei valori.
Basta un esempio, per capire meglio il funzionamento del test e la sua rilevanza. Si effettua l'esperimento classico di Mendel, andando a contare il numero delle piante che presentano determinate caratteristiche fenotipiche. I numeri dovrebbero essere nella proporzione 9:3:3:1 ma non lo sono perfettamente. Si deve accettare l'ipotesi che ha avanzato Mendel o la si può confutare?

Tipologia	N^o di casi osservati	N^o di casi aspettati
Lisci-gialli	315	312.75
Lisci-verdi	108	104.25
Rugosi-gialli	101	104.25
Rugosi-verdi	32	34.75

Si tengano in considerazione le seguenti informazioni. Supponiamo di avere in generale, n osservazioni raggruppate in k classi, A_1,A_2, . . . ,A_k; siano p_i le frequenze relative attese di ciascuna classe ($p_1 + p_2 + \ldots + p_k = 1$) e quindi np_1, np_2, . . . , np_k le frequenze assolute attese. Siano poi N_1, N_2, . . . ,N_k le frequenze assolute osservate. Calcoliamo in base a questi dati la seguente formula:

$$Q = \sum_{i=1}^{k} \frac{(np_i - N_i)^2}{np_i}.$$

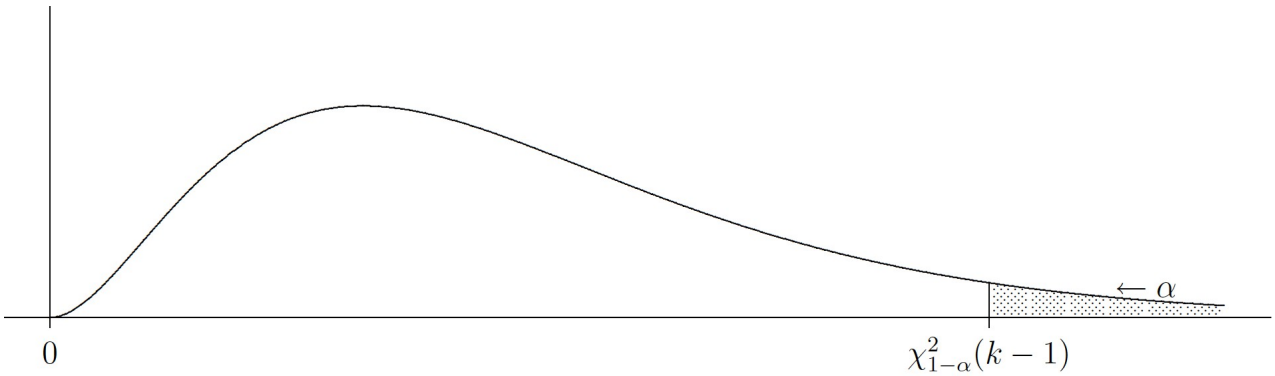
Si osservi che ogni addendo di Q ha a numeratore lo scarto quadratico tra le frequenza attesa e quella osservata, e a denominatore la frequenza attesa, che fa “pesare” diversamente i vari addendi. La *Q* sarà tanto più piccola quanto migliore è l’adattamento delle frequenze osservate a quelle attese. Inoltre la discrepanza tra frequenze osservate e attese è pesata più o meno a seconda della frequenza attesa. A parità di discrepanza pesa di più quella relativa a frequenze attese più piccole.
Se l’ipotesi nulla è
*H*0 : le osservazioni si adattano ai dati teorici,
il test sarà del tipo “Si rifiuti H_0 se $Q > k$” con *k* opportuno.
Il risultato fondamentale che permette di determinare il *k* opportuno è dato dal fatto che se *n è* grande allora *Q* ha una distribuzione che tende ad una legge $\chi^2(k-1)$, ovvero

$$\sum_{i=1}^{k} \frac{(np_i - N_i)^2}{np_i} \simeq W \sim \chi^2(k-1).$$

Questo permette di calcolare in modo completo la regione di rifiuto. La regione di rifiuto è l'area (α) a destra del quantile. Se vogliamo un test al livello α, la regione di rifiuto è

$$\{Q > \chi^2_{1-\alpha}(k-1)\}.$$

Quindi se Q si trova in α l'ipotesi è da rifiutare.
La condizione di applicabilità dell'approssimazione è che $np_i > 5$ per ogni i = 1, 2, . . . , k.



Nel caso esposto nell'esempio le frequenze attese sono tutte maggiori di 5, pertanto possiamo procedere al calcolo della quantità *Q*.

$$Q = \frac{(315-312.75)^2}{312.75} + \frac{(108-104.25)^2}{104.25} + \frac{(101-104.25)^2}{104.25} + \frac{(32-34.75)^2}{34.75} = 0.470.$$

Poiché ci sono 4 modalità, cioè 4 fenotipi diversi di piselli, il numero dei gradi di libertà è 3.
Per la scelta di α, convenzionalmente si scelgono valori come 0.01 o 0.05, perché l'ipotesi possa essere scientificamente plausibile.

$\chi^2_{.99}(3) = 11.3$, così che non si può rifiutare la teoria al livello dello 0.01;
$\chi^2_{.95}(3) = 7.81$, così che non si può rifiutare la teoria al livello dello 0.05.

Concludiamo che c'è una consistente probabilità che la distribuzione teorica concordi con l'esperimento.
La distribuzione χ^2 si presta anche ad altri test come, ad esempio, il test del χ^2 di indipendenza, che viene utilizzato per verificare l'indipendenza tra due variabili.