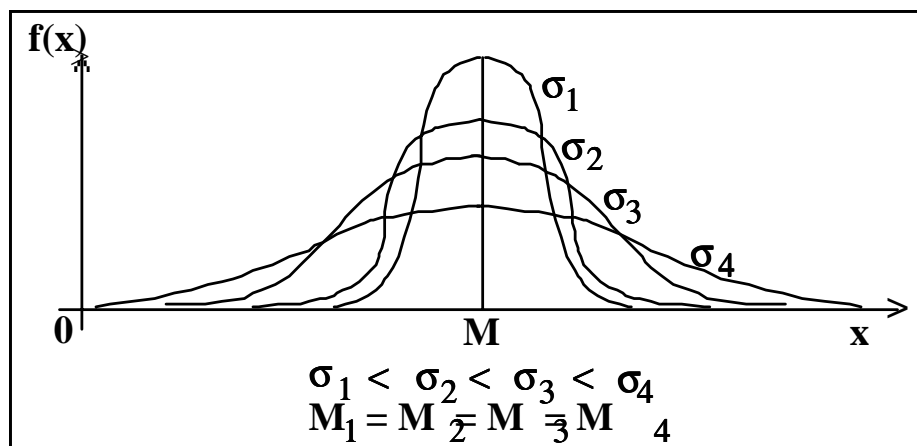


A parità di valore medio, la forma campanulare è tanto più appiattita quanto più grande è la variabilità



Date n variabili indipendenti x_1, x_2, \dots, x_n , se la loro distribuzione è di tipo normale, con media m_1, m_2, \dots, m_n e varianza $\sigma_{x_1}, \sigma_{x_2}, \dots, \sigma_{x_n}$ allora la variabile $X = x_1 + x_2 + \dots + x_n$, somma di queste variabili è anch'essa distribuita normalmente, con media $M = m_1 + m_2 + \dots + m_n$ e varianza $\Sigma = \sigma_{x_1} + \sigma_{x_2} + \dots + \sigma_{x_n}$, uguali alla somma delle medie ed alla somma delle varianze delle variabili originarie.

Momenti della distribuzione normale

La media e la deviazione standard della distribuzione normale sono uguali ad M ed a σ che compaiono nell'equazione normale.

Il momento di ordine k rispetto alla media è definito da:

$$\mu_k = \int_{-\infty}^{+\infty} \frac{(x - M)^k \cdot f(x)}{N} \cdot dx =$$

Data la simmetria della distribuzione tutti i momenti di ordine dispari rispetto alla media sono nulli

Per i momenti di ordine pari vale la relazione:

$$\mu_k = (k-1) \cdot \sigma^2 \cdot \mu_{k-2}.$$

per cui si avrà: $\mu_0 = 1, \mu_1 = 0, \mu_2 = \sigma^2, \mu_4 = 3 \cdot \sigma^4, \mu_6 = 15 \cdot \sigma^6$ e così via.

Il parametro σ della legge normale è dunque lo scarto tipo o deviazione standard di x

*La variabile t , trasformata della x che ha origine degli assi nel punto medio e unità di misura pari alla deviazione standard della variabile originaria, è detta anche *variabile normalizzata o scarto ridotto* e la sua legge di distribuzione viene detta *forma ridotta della legge normale*.*

Non dipende da alcun parametro, ha media uguale a zero e deviazione standard uguale all'unità. I momenti dei successivi ordini fino al quarto hanno i valori $\alpha_3 = 0$ ed $\alpha_4 = 3$ e lo stesso dicasi per β_1 e β_2 .

Questo risultato è molto importante, in quanto è a questi valori che vengono paragonati quelli corrispondenti calcolati per le altre distribuzioni di tipo campanulare, al fine di valutarne il grado di asimmetria e di appiattimento.

Adattamento della distribuzione normale a dati empirici

La densità di frequenza della distribuzione normale in x è :

$$f(x) = \frac{N}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{(x-M)^2}{2\sigma^2}}$$

che in termini di logaritmi diviene:

$$\log f(x) = -\frac{(x-M)^2}{2\sigma^2} + \log \frac{N}{\sigma\sqrt{2\pi}}$$

e ponendo $Y = \log f(x)$ ed $X = (x - M)^2$, esprimerà la retta:

$$Y = -\frac{1}{2\sigma^2} \cdot X + \log \frac{N}{\sigma\sqrt{2\pi}}$$

che può essere adattata ai dati effettivi con uno dei metodi di interpolazione disponibili.

Esempio IV.4 Si consideri la seguente tabella delle frequenze dell'altezza in centimetri di 78 piante di una determinata specie. Ci si propone di:

a) verificare se la curva normale può essere scelta come una buona curva interpolatrice per tale distribuzione di frequenze;

b) calcolare le frequenze teoriche della curva normale che ha la stessa media e la stessa varianza della seriazione data.

Altezza (cm)	Frequenza
10 - 15	4
15 - 20	20
20 - 25	28
25 - 30	12
30 - 35	8
35 - 40	6
	78

Calcoliamo anzitutto i momenti della distribuzione data.

Classi di altezza (cm)	Media di classe (v _i)	x _i =(v _i -22,5)					
		f _i	x _i f _i	x _i ² f _i	x _i ³ f _i	x _i ⁴ f _i	
10 - 15	12,5	4	-2	-8	16	-32	64
15 - 20	17,5	20	-1	-20	20	-20	20
20 - 25	22,5	28	0	0	0	0	0
25 - 30	27,5	12	1	12	12	12	12
30 - 35	32,5	8	2	16	32	64	128
35 - 40	37,5	6	3	18	54	162	486
Totali		78	3	18	134	186	710

$$M = 18/78 = 0,23; \quad M_2 = 134/78 = 1,72 ;$$

$$M_3 = 186/78 = 2,38; \quad M_4 = 710/78 = 9,10 ;$$

$$\mu_2 = M_2 - M^2 = 1,718 - 0,053 = 1,665 .$$

Con le correzioni di Sheppard sul momento secondo sarà:

$$\mu'_2 = \mu_2 - 1/12 = 1,66 - 0,08 = 1,58$$

$$\text{e quindi } \sigma = \text{radq}(1,58) = 1,26.$$

Per il momento terzo avremo:

$$\begin{aligned} \mu'_3 &= M_3 - 3 \cdot M_2 \cdot M + 2 \cdot M^3 = \\ &= 2,38 - 1,19 + 0,024 = 1,21 \end{aligned}$$

$$\alpha_3 = 1,21/(1,26)^3 = 1,21/1,99 = 0,6$$

e per il momento quarto μ_4 avremo:

$$\begin{aligned} \mu_4 &= M_4 - 4 \cdot M_3 \cdot M + 6 \cdot M_2 \cdot M^2 - 3 \cdot M^4 = \\ &= 9,10 - 2,19 + 0,55 - 0,01 = 7,44 \end{aligned}$$

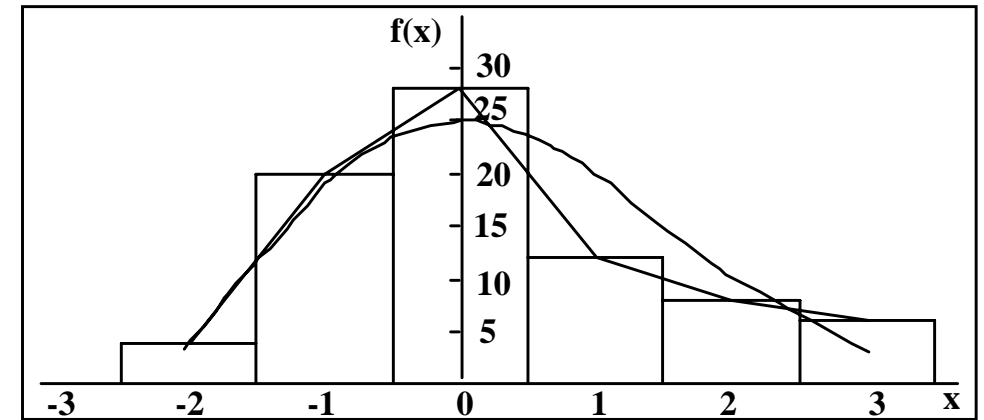
$$\begin{aligned}\mu'_4 &= \mu_4 - 1/2\mu_2 + 7/240 = \\ &= 7,44 - 0,83 + 0,03 = 6,64\end{aligned}$$

$$\alpha_4 = 6,64/(1,26)^4 = 6,64/2,50 = 2,6.$$

Come si vede, i valori di α_3 e α_4 sono quasi prossimi a zero ed a 3; si può quindi concludere che la curva normale può essere adottata come curva interpolatrice, malgrado essa non dia un perfetto adattamento ai dati.

La curva della distribuzione data ha *asimmetria positiva* e quindi essa sarà più allungata verso destra e più inclinata verso sinistra. Inoltre, poichè $\alpha_4 = 2,6$ è minore di 3, la curva sarà *più appiattita* della curva normale.

La rappresentazione grafica della distribuzione osservata e della curva normale ad essa adattata risulta essere la seguente:



Distribuzione rettangolare o uniforme

E' la più semplice tra le distribuzioni continue. Nell'intervallo tra $x_1=\alpha$ e $x_2=\beta$ ha densità di frequenza relativa pari a:

$$f(x) = \frac{1}{\beta - \alpha} \text{ con } (\alpha < x < \beta)$$

quindi ha *densità costante* in tutto l'intervallo compreso tra α e β . La rappresentazione grafica di questa distribuzione ha la forma di un rettangolo, che giustifica il suo nome.

E' l'equivalente della distribuzione uniforme discreta considerata nel continuo e la sua *media* e *varianza* sono:

$$M = \frac{\alpha + \beta}{2} \text{ e } \sigma^2 = \frac{(\beta - \alpha)^2}{12}$$

La mediana coincide con la media, mentre la moda o non esiste o ve ne sono tante quanti i valori compresi tra α e β .

Distribuzione esponenziale negativa

Distribuzione continua descritta dalla relazione

$$f(x) = \alpha \cdot e^{-\alpha x} \text{ con } \alpha > 0 \text{ e } x > 0$$

E' una funzione *positiva o nulla continuamente decrescente*, che tende a 0 per x *tendente all'infinito*. Nel discreto ha l'equivalente nella distribuzione geometrica decrescente.

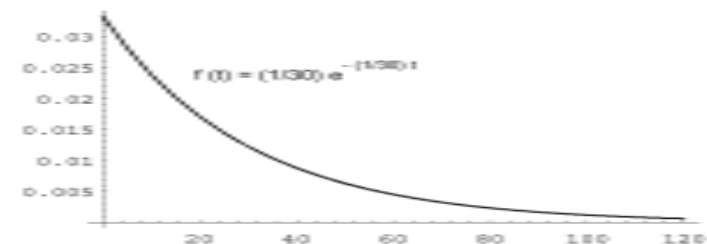


Figure 2.20: Exponential density with $\lambda = 1/30$.

La *media* e la *varianza* sono, rispettivamente:

$$\boxed{\mathbf{M} = \frac{1}{\alpha}} \text{ e } \boxed{\sigma^2 = \frac{1}{\alpha^2} = \mathbf{M}^2}.$$

Distribuzioni congiunte e condizionate

	Not smoke	Smoke	Total
Not cancer	40	10	50
Cancer	7	3	10
Totals	47	13	60

Table 4.1: Smoking and cancer.

		S	
		0	1
C	0	40/60	10/60
	1	7/60	3/60

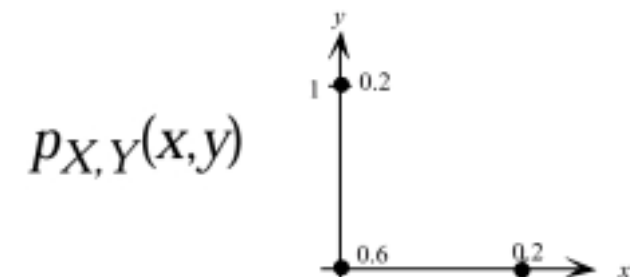
Table 4.2: Joint distribution.

$$\sum_x \sum_y p_{X,Y}(x,y) = 1,$$

$$\sum_y p_{X,Y}(x,y) = p_X(x),$$

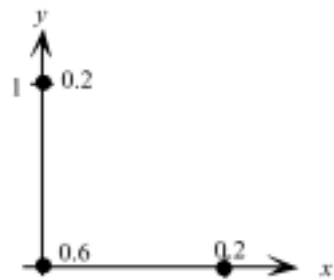
$$\sum_x p_{X,Y}(x,y) = p_Y(y).$$

ESEMPIO: In una scatola ci sono 5 lampadine di cui 4 buone ed una difettosa. Se ne estraggono 2. Sia x il numero di difetti della prima lampada ed y il numero di difetti della seconda lampada. Trovare la funzione di densità congiunta.



$$\begin{aligned} p_X(0) &= \sum_y p_{X,Y}(0,y) = p_{X,Y}(0,0) + p_{X,Y}(0,1) \\ &= 0.6 + 0.2 = 0.8 \end{aligned}$$

$$p_X(1) = \sum_y p_{X,Y}(1,y) = p_{X,Y}(1,0) = 0.2$$



$$p_X(x) = \begin{cases} 0.8 & \text{for } x = 0 \\ 0.2 & \text{for } x = 1 \\ 0 & \text{elsewhere} \end{cases}$$

$$p_Y(y) = \begin{cases} 0.8 & \text{for } y = 0 \\ 0.2 & \text{for } y = 1 \\ 0 & \text{elsewhere} \end{cases}$$

Distribuzioni condizionate

$$p_{X|Y}(x|y) = \frac{p_{X,Y}(x,y)}{p_Y(y)} \quad p_{Y|X}(y|x) = \frac{p_{X,Y}(x,y)}{p_X(x)}$$

$$p_{X,Y}(x,y) = p_X(x)p_{Y|X}(y|x) = p_Y(y)p_{X|Y}(x|y)$$

Variabili indipendenti

$$p_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n) = p_{X_1}(x_1)p_{X_2}(x_2) \cdots p_{X_n}(x_n)$$

X\Y	NON FUMO 0	FUMO 1	
NON Cancro 0	40/60	10/60	50/60
Cancro 1	7/60	3/60	10/60
	47/60	13/60	

$$p_{X|Y}(x|y) = \frac{p_{X,Y}(x,y)}{p_Y(y)}$$

$$P_{Y|X}(1|1) = \frac{3/60}{13/60} = \frac{3}{13}$$

Distribuzioni continue congiunte e condizionate

$$f(x|E) = \begin{cases} f(x)/P(E), & \text{if } x \in E, \\ 0, & \text{if } x \notin E. \end{cases}$$

$$P(F|E) = \int_F f(x|E) dx$$

$$P(F|E) = \int_F f(x|E) dx = \int_{E \cap F} \frac{f(x)}{P(E)} dx = \frac{P(E \cap F)}{P(E)}$$

$$\begin{aligned}
 G(t) &= \int_t^{\infty} \lambda e^{-\lambda x} dx \\
 &= -e^{-\lambda x} \Big|_t^{\infty} = e^{-\lambda t} \\
 P(F|E) &= \frac{P(F \cap E)}{P(E)} \\
 &= \frac{G(r+s)}{G(r)} \\
 &= \frac{e^{-\lambda(r+s)}}{e^{-\lambda r}} \\
 &= e^{-\lambda s} .
 \end{aligned}$$

Le curve di Karl Pearson

Il sistema di curve di frequenza di Karl Pearson riesce a descrivere con elevata approssimazione molte distribuzioni empiriche; si adatta molto bene a rappresentare la distribuzione dei valori in molti fenomeni reali, ma ha il grave limite che *i parametri che la definiscono non sono esplicativi*, spesso non forniscono alcun significato per l'interpretazione del fenomeno studiato e di conseguenza non si prestano ad usi predittivi.

La forma esplicita della funzione è espressa dall'equazione differenziale:

$$\frac{dy}{dx} = \frac{y(x+c)}{b_0 + b_1 \cdot x + b_2 \cdot x^2}$$

che dipende dalle radici dell'espressione quadratica del denominatore, cioè dai valori dei parametri b_0 , b_1 , b_2 , essendo x ed y i valori degli assi coordinati e c una costante.

Anche la distribuzione *normale* e le sue approssimazioni con *asimmetria* ed *appiattimento* variabili possono essere rappresentate come una delle possibili curve del Pearson e possono essere riguardate come una parte delle distribuzioni che è possibile descrivere con elevata precisione.

Distribuzione Gamma (III tipo del Pearson)

La distribuzione Gamma (Γ) descrive una variabile casuale *continua e positiva* distribuita con densità pari a

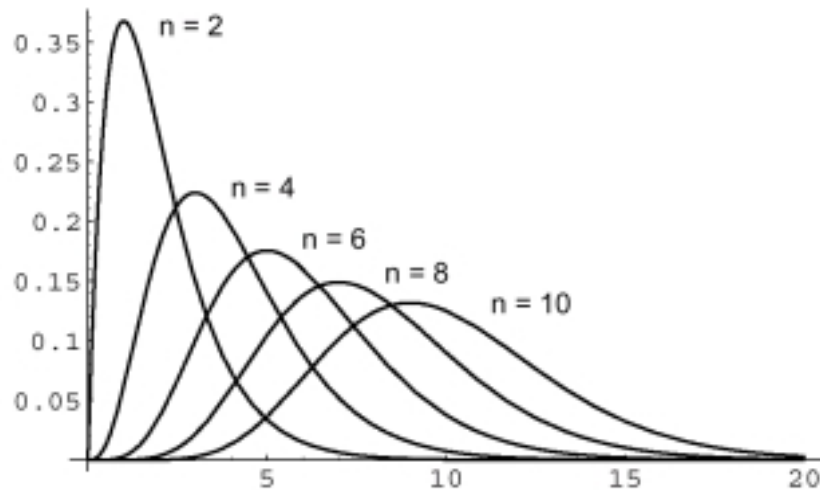
$$f(x) = \frac{1}{\Gamma(m)} x^{m-1} e^{-x} \quad (0 \leq x < \infty; m > 0)$$

in cui la funzione gamma è definita da:

$$\Gamma(m) = \int_0^{\infty} x^{m-1} e^{-x} dx$$

La funzione Gamma dipende dal parametro m e la variabile che descrive si chiama una variabile $\gamma(m)$.

Tale distribuzione è *asintotica* rispetto all'asse delle x e si annulla per $x=0$ se $n>2$.



Essa fu chiamata da Legendre, agli inizi dell'800, *integrale euleriano di seconda specie*, in onore di Eulero che la studiò per primo nella seconda metà del '700, e costituisce una curva del III tipo del Pearson.

La distribuzione Gamma è di notevole importanza in Statistica poiché si dimostra che:

- *quando una variabile statistica è distribuita normalmente, il suo quadrato è distribuito secondo una distribuzione Gamma.*
- *Inoltre, anche la distribuzione della varianza dei campioni è ancora una distribuzione gamma*
- *e, in generale, la somma dei quadrati di variabili normali è distribuita secondo una distribuzione Gamma.*

Distribuzione Beta (I tipo del Pearson)

Una variabile continua x di densità:

$$f(x) = \frac{x^{k-1}(1-x)^{m-1}}{B(k,m)} \quad (0 \leq x \leq 1; k > 0; m > 0) \text{ in cui}$$

$$B(k,m) = \frac{\Gamma(k) \cdot \Gamma(m)}{\Gamma(k+m)} = \int_0^1 x^{k-1} (1-x)^{m-1} dx$$

si dice distribuita secondo la funzione Beta con parametri k ed m o più brevemente chiamasi variabile $\beta(k,m)$. La suddetta $f(x)$ è una curva del *I tipo del Pearson*.

La distribuzione Beta è strettamente collegata con la distribuzione Gamma. Si dimostra che:

Se due variabili indipendenti sono distribuite rispettivamente secondo una $\chi(k)$ e una $\chi(m)$, il loro rapporto è distribuito secondo una $\beta(k,m)$.

Ciò è importante per la distribuzione F, fondamentale per l'analisi della varianza, che come rapporto di variabili distribuite secondo una Gamma, è una Beta.

Distribuzione Chi-quadrato $\chi^2(n)$

Se **Z1,Z2,Z3,...,Zn** sono variabili normali standardizzate ($\mu=0,\sigma=1$) indipendenti la variabile:

$$\chi_n^2 = \sum_{i=1}^n Z_i^2$$

è denominata variabile χ^2 con **n** gradi di libertà.
una variabile χ_n^2 è anche quindi:

$$\chi_n^2 = \sum_{i=1}^n \left(\frac{X_i - \mu_i}{\sigma_i} \right)^2$$

Distribuzione Chi-quadrato $\chi^2(n)$

$$f(\chi_n^2) = \frac{1}{\Gamma\left(\frac{n}{2}\right)} \cdot \left(\frac{1}{2}\right)^{n/2} \cdot (\chi^2)^{n/2-1} \cdot e^{-\frac{1}{2}\chi^2}$$

$$0 < \chi^2 < \infty$$

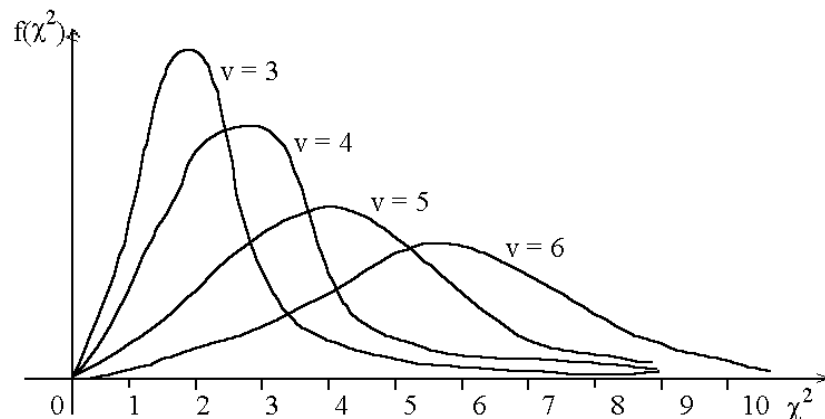
n rappresenta il numero dei *gradi di libertà*.

Essa è un caso particolare della distribuzione Gamma.

TUTTE LE SUE CARATTERISTICHE DIPENDONO SOLO DAL PARAMETRO **n** (NUMERO DEI GRADI DI LIBERTÀ)

PROPRIETÀ' DELLA VARIABILE χ^2

- ◆ **VARIABILE CONTINUA**
- ◆ **ASSUME VALORE POSITIVO FRA 0 ED INFINITO**



◆ LA VARIABILE SOMMA DI DUE VARIABILI χ^2 RISPETTIVAMENTE CON **N1** E **N2** GRADI DI LIBERTÀ' E' A SUA VOLTA DISTRIBUITA COME UNA χ^2 CON **N1+N2** GRADI DI LIBERTÀ'

◆ LA **MEDIA** DI UNA V.A. χ_n^2 è PARI AD **n**

◆ LA **VARIANZA** DI UNA V.A. χ_n^2 è PARI A **2n**

◆ LA DISTRIBUZIONE PER **N>100** HA UN ANDAMENTO CHE PUO' ESSERE APPROSSIMATO A QUELLO DI UNA VARIABILE NORMALE. NELLE TAVOLE NON SI TROVA QUINDI LA DISTRIBUZIONE χ_n^2 PER **N>100** IN QUANTO E' POSSIBILE RICAVARLA DALLE TAVOLE DELLA DISTRIBUZIONE NORMALE

♦ LA DISTRIBUZIONE χ_n^2 NON E' ADATTA A RAPPRESENTARE LA DISTRIBUZIONE DI POPOLAZIONI O L'ANDAMENTO DI FENOMENI IN GENERE.

♦ E' UTILE PER L'ANALISI DI STATISTICHE CAMPIONARIE

CALCOLO DELLA **MEDIA** DI UNA χ_n^2

$$E[\chi_n^2] = E\left[\sum_{i=1}^n \chi_i^2\right] = \sum_{i=1}^n E\left[\left(\frac{X_i - \mu_i}{\sigma_i}\right)^2\right] = \sum_{i=1}^n 1 = n$$

VARIANZA DI UNA χ_n^2

$$VAR(\chi_n^2) = E[(\chi_n^2)^2] - [E[\chi_n^2]]^2 = 2n$$

SOMMA DI DUE VARIABILI χ_n^2

$$K = \chi_{n1}^2 + \chi_{n2}^2 = \sum_{i=1}^{n1} \left(\frac{X_i - \mu_i}{\sigma_i}\right)^2 + \sum_{i=1}^{n2} \left(\frac{X_i - \mu_i}{\sigma_i}\right)^2 = \sum_{i=1}^{n1+n2} \left(\frac{X_i - \mu_i}{\sigma_i}\right)^2 \cong \chi_{n1+n2}^2$$

VARIANZA DEL CAMPIONE DI UNA DISTRIBUZIONE NORMALE

Per stimare la varianza σ^2 di una distribuzione normale si può usare lo stimatore :

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

per analizzare la distribuzione di questo stimatore è necessario utilizzare la funzione di distribuzione del **CHI-QUADRATO** si ha:

$$\left(\frac{n-1}{\sigma^2}\right) S^2 \approx \chi_{n-1}^2$$

RELAZIONE FRA UNA VARIABILE χ_n^2 ED UNA VARIABILE NORMALE PER N GRANDE

si ha che la quantita'

$$\sqrt{2\chi_n^2} - \sqrt{2n-1} \cong Z$$

SEGUE UNA LEGGE N(0,1) QUINDI:

$$\chi_n^2 \cong \frac{1}{2} (Z + \sqrt{2N-1})^2$$

DA CUI

$$P(\chi_n^2 < K) \approx P\left(\frac{1}{2} (Z + \sqrt{2N-1})^2 < K\right) = P(Z < \sqrt{2K} - \sqrt{2n-1})$$

E QUINDI

$$\chi_{n,\alpha}^2 \cong \frac{1}{2} (Z_\alpha + \sqrt{2N-1})^2$$

Distribuzione F di Fisher

La *distribuzione F* è una distribuzione sulla quale si basa molta parte della statistica parametrica. Essa corrisponde alla distribuzione del rapporto di due variabili casuali chi-quadrato indipendenti (A e B), divise per i loro rispettivi

gradi di libertà (m ed n), cioè:

$$F(m,n) = \frac{\frac{A}{m}}{\frac{B}{n}}$$

L'ordine con il quale sono riportati i due numeri che indicano i gradi di libertà di F è importante, perché la densità della distribuzione di F non è simmetrica rispetto ad essi. Per primo si riporta sempre il numero di gradi di libertà del numeratore e per secondo quello del denominatore.

Distribuzione t di Student

La *distribuzione t di Student*, con n gradi di libertà è data dal rapporto tra una variabile casuale con distribuzione normale standardizzata (Z) e la radice quadrata di una variabile casuale con distribuzione chi-quadrato (A), indipendente dalla Z e divisa per i suoi gradi di libertà (n), cioè:

$$t_{(n)} = \frac{Z}{\sqrt{\frac{A}{n}}}.$$

All'aumentare dei gradi di libertà, la distribuzione t di Student si avvicina alla distribuzione normale standardizzata.

Il quadrato di una t di Student con n gradi di libertà è uguale ad una distribuzione F di Fisher con gradi di libertà 1 ed n, cioè:

$$t_{(n)}^2 = F_{(1,n)},$$

oppure

$$t_{(n)} = \sqrt{F_{(1,n)}}.$$