Distribuzioni χ^2

Come nel caso delle distribuzioni t di Student, anche le distribuzioni χ^2 formano una famiglia di funzioni di densità parametrizzate dai gradi di libertà.

- C'è una diversa distribuzione χ^2 a seconda dei numeri di grado di libertà $\nu = 1, 2, \ldots$.
- Diversamente dalle distribuzioni t di Student e dalla distribuzione normale, la distribuzione χ^2 possiede un'asimmetria positiva.
- L'asimmetria della distribuzione diminuisce al crescere dei gradi di libertà.

Si chiama χ^2 la sommatoria dei quadrati di n variabili (casuali e indipendenti) normali standardizzate, che è espressa dalla seguente equazione:

$$\chi^2_n = \sum_{i=1}^{n} z_i^2 = \sum_{i=1}^{n} \left(\frac{x_i - \mu_i}{\sigma_i} \right)^2$$

- Dato che la variabile casuale χ^2 è generata dalla somma dei quadrati di n *valori indipendenti* di una variabile normale standardizzata, i gradi di libertà coincideranno con: $\nu = n$.
- Quando tali valori *non sono indipendenti*, è necessario stabilire le condizioni che li vincolano. Sottraendo tali vincoli si ottiene il numero di gradi di libertà.
- I gradi di libertà sono dunque il parametro che caratterizza ogni distribuzione di χ^2 e vengono molto spesso indicati con la lettera greca ν.

Approfondimento: i momenti caratteristici di una distribuzione

1. Si definisce momento di ordine k la quantità:

$$E\left(x_i - \mu\right)^k.$$

Il momento di ordine $k = 0$ è 1, il momento di ordine $k = 1$ è $E\left(x_i - \mu\right) = 0$, infine il momento di ordine $k = 2$ è la varianza $\sigma^2 = E\left(x_i - \mu\right)^2$.

2. Per la *distribuzione normale*, possiamo esprimere i momenti di ordine $k > 1$ in funzione di σ^2 attraverso la seguente formula generale:

$$(|k-1|)\sigma^2 E\left(x - \mu\right)^{|k-2|}$$

da cui facilmente ricaviamo che:

$$\begin{array}{lll}
k = 1; & (|1-1|)\sigma^2 E\left(x-\mu\right)^{|1-2|} & = 0 \times \sigma^2 \times 0 \quad = 0 \\
k = 2; & (|2-1|)\sigma^2 E\left(x-\mu\right)^{|2-2|} & = 1 \times \sigma^2 \times 1 \quad = \sigma^2 \\
k = 3; & (|3-1|)\sigma^2 E\left(x-\mu\right)^{|3-2|} & = 2\sigma^2 \times 0 \quad = 0 \\
k = 4; & (|4-1|)\sigma^2 E\left(x-\mu\right)^{|4-2|} & = 3\sigma^2\sigma^2 \quad = 3(\sigma^2)^2
\end{array}$$

3. Si definisce l'indice di asimmetria come il rapporto tra il momento di ordine $k = 3$ e il cubo della deviazione standard:

$$\text{asimmetria} = \frac{E\left(x-\mu\right)^3}{\sigma^3} = E\left[\frac{(x-\mu)}{\sigma}\right]^3,$$

Detto anche *momento standardizzato di ordine terzo*, assume valore 0 nella distribuzione normale standard (vedi punto 2; per $k = 3$)

4. L'indice di curtosi, o *momento standardizzato di ordine quarto*:

$$\text{kurtosi} = \frac{E\left(x-\mu\right)^4}{(\sigma^2)^2} = E\left[\frac{(x-\mu)}{\sigma}\right]^4,$$

assume valore 3 nella distribuzione normale e quantifica l'appiattimento di una distribuzione. Le distribuzioni piatte con code ampie sono chiamate *platicurtiche* (es. $t - Student$), quelle appuntite con code piccole sono chiamate *leptocurtiche*. Una distribuzione con la stessa kurtosi della distribuzione normale è chiamata *mesocurtica*. (vedi punto 2; per $k = 4$)

Premesse

Si dimostrano le seguenti equivalenze asintotiche (per $n \to \infty$):

5 Il valore atteso di un valore $z^2 \sim \chi^2_{\nu=1}$ è uguale al valore 1:

$$E\left[\chi^2_1\right] = E\left[\frac{(x_i - \mu)^2}{\sigma^2}\right] = \left[\frac{1}{\sigma^2} E\left(x_i - \mu\right)^2\right] = \left[\frac{1}{\sigma^2}\ \sigma^2\right] = 1$$

6 Il valore atteso di un valore $z^4 \sim (\chi^2_{\nu=1})^2$ è uguale al valore 3:

$$E\left[\left(\chi^2_1\right)^2\right] = E\left[\frac{(x_i - \mu)^4}{(\sigma^2)^2}\right] = \left[\frac{E\left(x_i - \mu\right)^4}{(\sigma^2)^2}\right] = \text{indice di curtosi} = 3$$

7 La somma di n valori $\chi^2_{\nu=1}$ è un $\chi^2_{\nu=n}$:

$$\sum_{i=1}^{n} \left(\chi^2_1\right)_i = \sum_{i=1}^{n} \left(\frac{(x_i - \mu)^2}{\sigma^2}\right) = \sum_{i=1}^{n} z_i^2 = \chi^2_n$$

Distribuzioni χ^2 - valore atteso

Si dimostra che:

Il valore atteso di una variabile χ^2_n è uguale al numero dei gradi di libertà della variabile stessa:

$$E\left(\chi^2_n\right) = n = \nu$$

Applicazione dei punti 5. e 7.

$$E\left[\chi^2_n\right] = E\left[\sum_{i=1}^{n} \left(\chi^2_1\right)_i\right] = \sum_{i=1}^{n} \left[E\left(\chi^2_1\right)_i\right] = n[1] = n$$

Distribuzioni χ^2 - varianza

la varianza è pari a due volte i gradi di libertà:

$$Var\left(\chi^2_n\right) = 2n = 2\nu$$

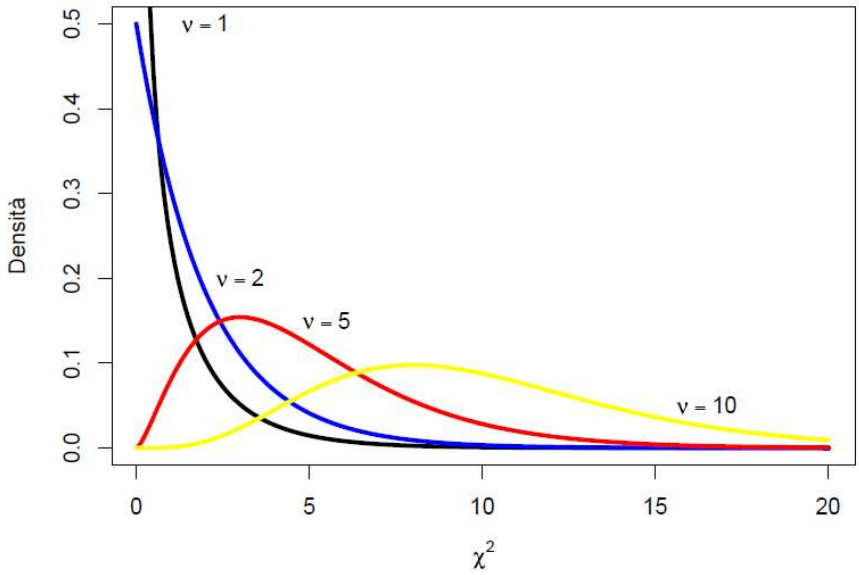
Applicazione dei punti 5. 6. e 7.

$$Var\left[\chi^2_n\right] = Var\left[\sum_{i=1}^{n}\left(\chi^2_1\right)_i\right] = \sum_{i=1}^{n}\left[Var\left(\chi^2_1\right)_i\right] =$$

$$= \sum_{i=1}^{n}\left[E\left[\left(\chi^2_1\right)^2_i\right] - \left[E\left(\chi^2_1\right)_i\right]^2\right] =$$

$$= n\left[3 - 1^2\right] = 2n$$

- La distribuzione χ^2 è di tipo continuo e non può mai essere negativa; perciò si trova sempre compresa nel primo quadrante degli assi cartesiani ed ha forme diverse a seconda del valore ν.



- La variabile χ^2 è definita nell'intervallo $[0; \infty]$.

Distribuzione campionaria di s^2

- Possiamo utilizzare la distribuzione χ^2 per ricavare la distribuzione campionaria della statistica

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

- utilizzeremo tale informazione per fare inferenze sulla varianza incognita della popolazione.
- In particolare, svilupperemo la stima per intervalli di confidenza del parametro σ, a partire dal valore campionario s^2.
- Concentriamoci sul numeratore

$$\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- che possiamo esprimere come sottrazione di μ:

$$\sum_{i=1}^{n}\left[(x_i - \mu) - (\bar{x} - \mu)\right]^2;$$

- e da cui sviluppando il quadrato, otteniamo:

$$s^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n}(x_i - \mu)^2 - n(\bar{x} - \mu)^2\right].$$

$$\sum_{i=1}^{n}\left[(x_i-\mu)-(\bar{x}-\mu)\right]^2 = \sum_{i=1}^{n}(x_i-\mu)^2 + n(\bar{x}-\mu)^2 - 2(\bar{x}-\mu)\sum_{i=1}^{n}(x_i-\mu)$$

$$= \sum_{i=1}^{n}(x_i-\mu)^2 + n(\bar{x}-\mu)^2 - 2(\bar{x}-\mu)(n\bar{x}-n\mu)$$

$$= \sum_{i=1}^{n}(x_i-\mu)^2 + n(\bar{x}-\mu)^2 - 2n(\bar{x}-\mu)^2$$

$$= \sum_{i=1}^{n}(x_i-\mu)^2 - n(\bar{x}-\mu)^2$$

Quindi

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left[(x_i-\bar{x})\right]^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n}(x_i-\mu)^2 - n(\bar{x}-\mu)^2\right]$$

- Moltiplicando entrambi i termini dell'equazione per la costante $\frac{n-1}{\sigma^2}$ si ottiene

$$\left(\frac{n-1}{\sigma^2}\right)s^2 = \left(\frac{n-1}{\sigma^2}\right)\frac{1}{n-1}\left[\sum_{i=1}^{n}(x_i-\mu)^2 - n(\bar{x}-\mu)^2\right]$$

$$= \sum_{i=1}^{n}\frac{(x_i-\mu)^2}{\sigma^2} - \frac{(\bar{x}-\mu)^2}{\sigma^2/n}$$

$$= \sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma}\right)^2 - \left(\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}\right)^2$$

$$= \sum_{i=1}^{n} z_i^2 - z_{\bar{x}}^2 \sim \chi^2_{\nu=n-1}$$

- dunque la quantità $\left(\frac{n-1}{\sigma^2}\right) s^2$ corrisponde ad una sommatoria di variabili normali standardizzate, che approssima una distribuzione χ^2 con $\nu = n - 1$ gradi di libertà.
- Si noti infatti che $\frac{x_i - \mu}{\sigma}$ è una variabile (*normale*) standardizzata, con $\nu_1 = n$ gradi di libertà (μ non è stimata), mentre $\frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$ è la media campionaria standardizzata, con $\nu_2 = 1$ grado di libertà (ancora, μ non è stimata);
- allora, la loro sottrazione si distribuisce secondo la legge $\chi^2_{\nu_1} - \chi^2_{\nu_2} = \chi^2_{\nu_1 - \nu_2} = \chi^2_{n-1}$

$$\left(\frac{n-1}{\sigma^2}\right) s^2 \sim \chi^2_{n-1}.$$

Valore atteso e varianza di s^2

- Possiamo ricavare facilmente il valore atteso e la varianza di s^2 sfruttando l'approssimazione $\left(\frac{n-1}{\sigma^2}\right) s^2 \sim \chi^2_{n-1}$:

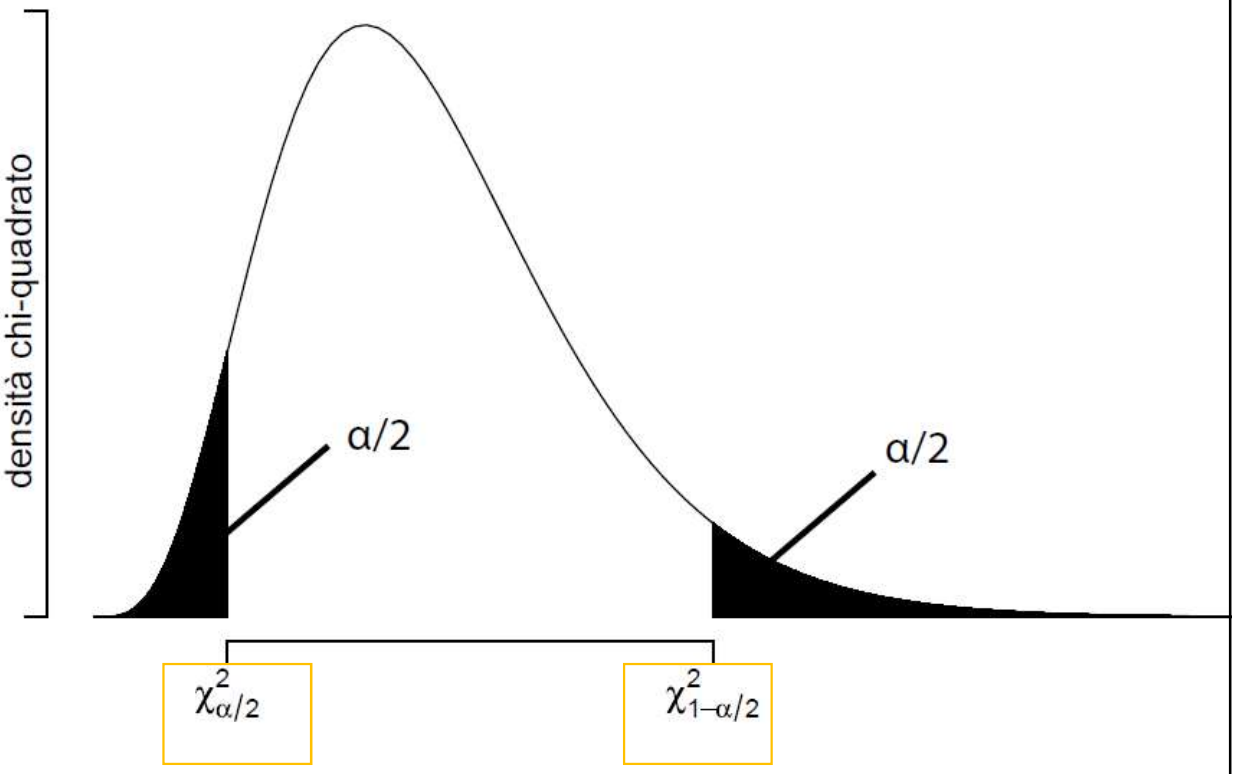
$$E(s^2) = E\left(\frac{n-1}{\sigma^2} s^2 \frac{\sigma^2}{n-1}\right) = E\left(\chi^2_{n-1} \frac{\sigma^2}{n-1}\right) = E\left(\chi^2_{n-1}\right) \frac{\sigma^2}{n-1} = \sigma^2$$

$$Var\left(s^2\right) = Var\left(\chi^2_{n-1} \frac{\sigma^2}{n-1}\right) = Var\left(\chi^2_{n-1}\right) \frac{\left(\sigma^2\right)^2}{(n-1)^2} =$$

$$= 2\left(n-1\right) \frac{\left(\sigma^2\right)^2}{(n-1)^2} = \frac{2\sigma^4}{(n-1)}$$

Stima per intervalli di σ^2

- In certi casi può essere necessario fare un'inferenza sulla varianza della popolazione σ^2.
- Abbiamo detto che $s^2 = \sum (x - \bar{x})^2/(n-1)$ è uno stimatore corretto di σ, e quindi serve naturalmente a tale scopo.
- Sappiamo inoltre che
  - avendo a disposizione un campione di n valori indipendenti $x_1, x_2, \ldots, x_n$;
  - calcolando la statistica $(n-1)\, s^2/\sigma^2$;
  - dalle tavole della distribuzione χ^2_{n-1}, per un coefficiente di fiducia α, si possono ottenere i valori $\chi^2_{(\alpha/2)}$ e $\chi^2_{(1-\alpha/2)}$ atti a delimitare un intervallo centrale contenente una probabilità di $1-\alpha$.



N.B Diversa scrittura rispetto al testo di Luccio!
Cambia solamente la posizione nell'intervallo.

Dato che la distribuzione χ^2 è asimmetrica, almeno per n non troppo elevati, la centralità dell'intervallo va intesa in senso probabilistico, avendo posto $\alpha/2$ di probabilità su ogni coda:

$$P\left(\chi^2_{(\alpha/2)} \leq \frac{(n-1)s^2}{\sigma^2} \leq \chi^2_{1-\alpha/2}\right) = 1-\alpha$$

prendendo i reciproci e quindi invertendo la diseguaglianza

$$P\left(\frac{1}{\chi^2_{(\alpha/2)}} \geq \frac{\sigma^2}{(n-1)s^2} \geq \frac{1}{\chi^2_{1-\alpha/2}}\right) = 1-\alpha$$

e moltiplicando gli estremi per $(n-1)s^2$

$$P\left(\frac{(n-1)s^2}{\chi^2_{(\alpha/2)}} \geq \sigma^2 \geq \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}\right) = 1-\alpha$$

Illustrazione. Si supponga di aver scelto a caso 16 scolaresche dalla popolazione italiana, omogenee come numero, sesso ed età dei componenti. Per queste scolaresche si rileva il tempo dedicato ad attività ricreative. Le statistiche riassuntive del campione danno una media pari a 3.3375, una varianza $s^2 = 2.897$ e $s = 1.702$.

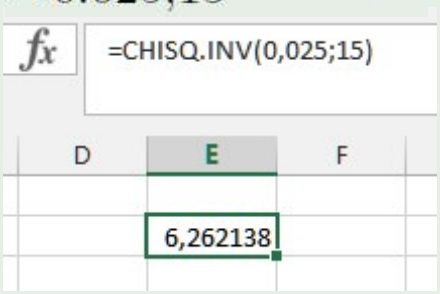
Si calcoli un intervallo di confidenza al 95% per σ^2.

Si userà la varianza campionaria $s^2 = 2.897$ come stimatore non distorto di σ^2 e quindi ci si servirà dei valori

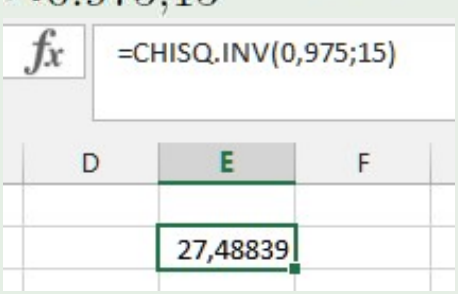
$$\frac{(n-1)s^2}{\chi^2_{0.975}} \quad \text{e} \quad \frac{(n-1)s^2}{\chi^2_{0.025}}$$

rispettivamente, come estremo inferiore e superiore dell'intervallo di fiducia.

$$\chi^2_{0.025;15} = 6.26$$



$$\chi^2_{0.975;15} = 27.49$$



quindi l'intervallo di fiducia per σ^2 al 95% sarà:

$$P\left(\frac{(15)2.897}{6.26} \geq \sigma^2 \geq \frac{(15)2.897}{27.49}\right) = 1 - \alpha$$

$$P\left(6.94 \geq \sigma^2 \geq 1.58\right) = 1 - \alpha$$