STATISTICA E ANALISI DEI DATI

Capitolo 14 – Verifica delle Ipotesi

Dott. Stefano Cirillo
Dott. Luigi Di Biasi

a.a. 2025-2026

Criterio del chi-quadrato

- Ci siamo finora occupati di ricavare informazioni da un campione estratto da una popolazione descritta da una variabile aleatoria X caratterizzata da una funzione di probabilità o densità di probabilità $f(x;\vartheta)$
  - stimando il parametro/i non noto/i ϑ della popolazione con stime puntuali ed intervallari
- Vogliamo ora verificare se il campione osservato può **essere stato estratto** da una popolazione descritta da una variabile aleatoria X con funzione di distribuzione $F_X(x)$
  - Utilizzeremo il criterio di verifica **delle ipotesi del chi-quadrato**, detto anche test del chi-quadrato o test del buon adattamento

Criterio del chi-quadrato bilaterale

- Vogliamo verificare l'ipotesi che un certa popolazione, descritta da una variabile aleatoria X, **sia caratterizzata da una funzione di distribuzione** $F_X(x)$, con k parametri **non noti** da stimare
- Denotiamo con:
  - $\mathbf{H_0}$ l'**ipotesi nulla** soggetta a verifica
  - $\mathbf{H_1}$ l'**ipotesi alternativa**

  il test chi-quadrato con livello di significatività α mira a verificare:
  - $\mathbf{H_0}$: X ha una funzione di distribuzione $F_X(x)$
  - $\mathbf{H_1}$: X non ha una funzione di distribuzione $F_X(x)$

  considerando α come la probabilità massima di rifiutare l'ipotesi nulla quando essa è vera
- Occorre **determinare un test** con livello di significatività α che permetta di determinare una regione di accettazione e di rifiuto dell'ipotesi nulla
  - Il test di verifica delle ipotesi considerato è bilaterale (o a due code)

Criterio del chi-quadrato bilaterale

- Suddividiamo l'insieme dei valori che la variabile aleatoria X può assumere in r **sottoinsiemi / intervalli** $I_1, I_2, \dots, I_r$ in modo che la probabilità p_i che la variabile aleatoria assuma un valore appartenente I_i:

$$p_i \in P(X \in I_i)$$

- Estraiamo un campione $x_1, x_2, \dots, x_n$ di ampiezza n
  - Osserviamo le **frequenze assolute** del campione $n_1, n_2, \dots, n_r$ con cui gli n elementi si **distribuiscono** nei rispettivi insiemi $I_1, I_2, \dots, I_r$
- Si ha quindi che:

$$p_i \geq 0 \; con \boxed{\sum_{i=1}^{r} p_i = 1} \rightarrow$$

La somma delle probabilità che X assuma un valore appartenente I_i è 1

$$n_i \geq 0 \; con \boxed{\sum_{i=1}^{n} n_i = n} \rightarrow$$

La somma delle frequenze con cui gli elementi si distribuiscono nei rispettivi insiemi è uguale alla grandezza del campione n

$$(i = 1, 2, \dots, r)$$

Selezione degli Intervalli

- Gli intervalli permettono di suddividere il dominio della variabile casuale X in un numero finito di classi o categorie

- Questi intervalli sono fondamentali per **calcolare le frequenze osservate** e attese e verificare se la distribuzione empirica di X è coerente con la distribuzione teorica $F_X(x)$
  - Si suddivide il dominio di X in r intervalli disgiunti $I_1, I_2, \dots, I_r$, che coprono tutto il dominio della variabile
  - Gli estremi degli intervalli possono essere scelti in modo equidistante o meno, a seconda della distribuzione di $F_X(x)$
  - Gli intervalli devono essere definiti in modo tale **che il numero di osservazioni attese in ciascun intervallo** sia sufficientemente grande, **idealmente maggiore o uguale a 5**
    - Questo assicura che l'approssimazione del chi quadrato sia valida.

$$\min(np_1, np_2, \dots, np_r) \geq 5$$

Criterio del chi-quadrato bilaterale

- Si nota che la probabilità:

Probabilità che esattamente n_1 elementi appartengano ad I_1, n_2 elementi appartengano ad I_2, …, n_r elementi appartengano ad I_r

$$\boxed{p(n_1, n_2, \dots, n_r)} = \frac{n!}{n_1! n_2! \dots n_r!} p_1^{n_1} p_2^{n_2} \dots p_r^{n_r}$$

è la probabilità di una specifica distribuzione dei conteggi $n_1, n_2, \dots, n_r$ tra gli r intervalli $I_1, I_2, \dots, I_r$ ed è basata sulla distribuzione multinomiale

- **Approssimazione normale per grandi campioni**
  - Per n grande, per il **teorema del limite centrale**:
    - Ogni N_i ha media np_i e varianza $np_i(1 - p_i)$.
    - I termini non sono indipendenti perché $\sum N_i = n$.
    - La distribuzione congiunta di $(N_1, \dots, N_r)$ è approssimativamente normale
- Si calcola poi la quantità:

Numero medio di elementi che cadono nell'intervallo I_i è np_i

$$\chi^2 = \sum_{i=1}^{r} \left(\frac{n_i - \boxed{np_i}}{\sqrt{np_i}} \right)^2$$

Criterio del chi-quadrato bilaterale

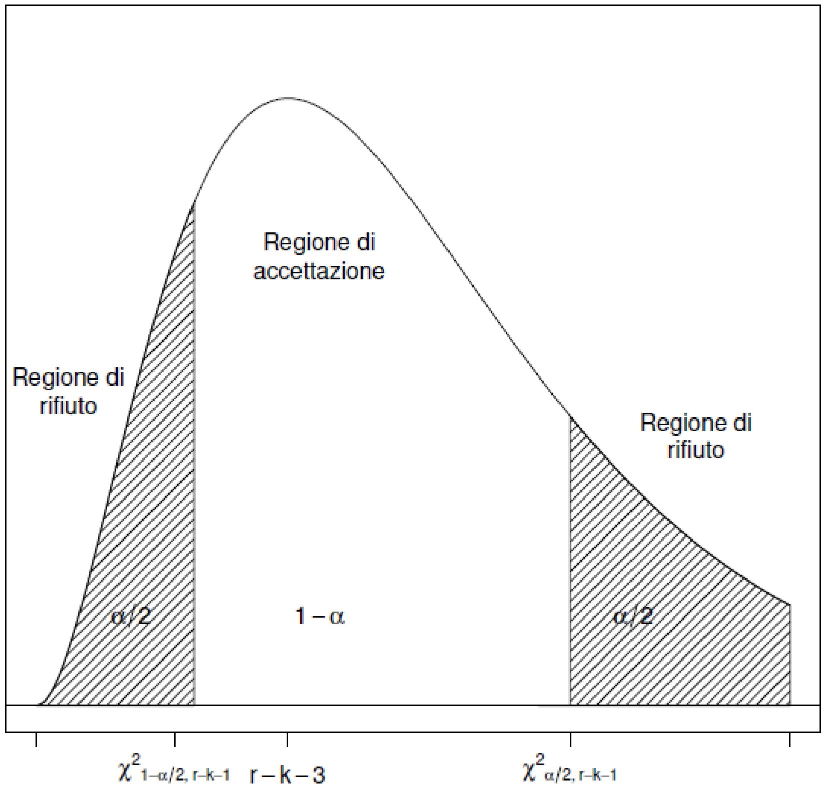
- Se la variabile aleatoria X ha una funzione di distribuzione $F_X(x)$ con k parametri non noti
  - Il criterio chi-quadrato si basa sulla statistica:

$$Q = \sum_{i=1}^{r} \left(\frac{N_i - np_i}{\sqrt{np_i}} \right)^2$$

Che è **approssimabile con la funzione di distribuzione chi-quadrato** con $r - k - 1$ gradi di libertà

- Si sottrae:
  - 1 da r in quanto se conosciamo $r - 1$ delle probabilità p_i la rimanente probabilità può essere univocamente determinata
  - k poiché si suppone che siano k i parametri indipendenti non noti sostituiti da stimare

Densità chi–quadrato con r–k–1 gradi di libertà



Criterio del chi-quadrato bilaterale

- **Definizione**:
  - Per un campione **sufficientemente** numeroso di ampiezza n, il test chi-quadrato bilaterale di misura α è il seguente:
    - Ipotesi $\mathbf{H_0}$: X ha una funzione di distribuzione $F_X(x)$
      - Si **accetti** l'ipotesi $\mathbf{H_0}$ se:

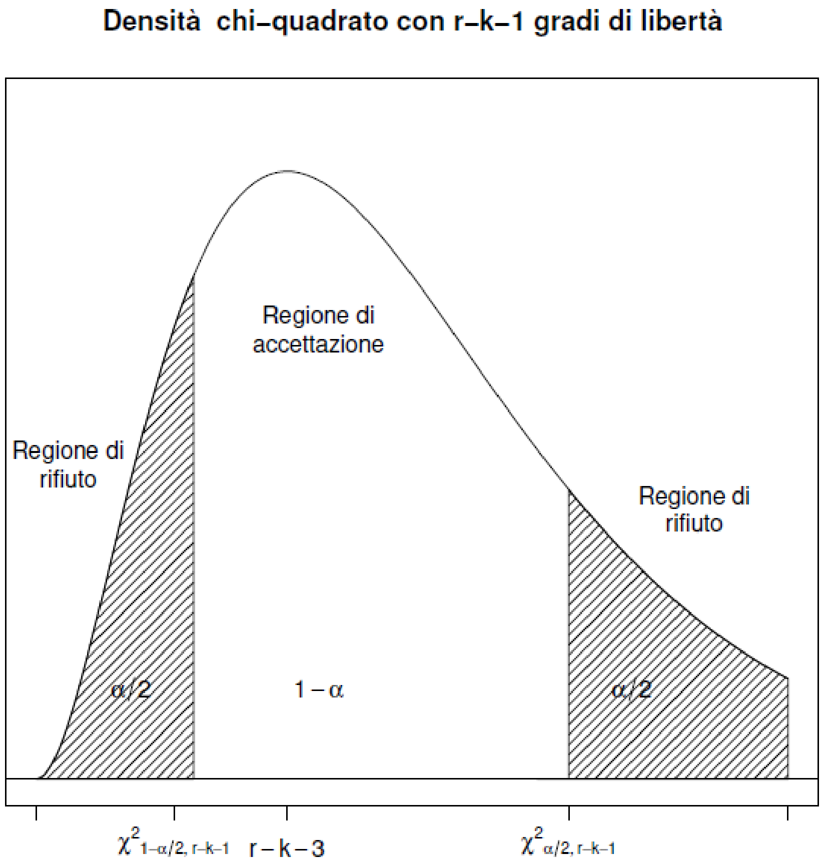
$$\chi^2_{1-\alpha/2,r-k-1} < \chi^2 < \chi^2_{\alpha/2,r-k-1}$$

      - Si **rifiuti** l'ipotesi $\mathbf{H_0}$ se:

$$\chi^2 < \chi^2_{1-\alpha/2,r-k-1} \ oppure \ \chi^2 > \chi^2_{\alpha/2,r-k-1}$$

dove $\chi^2_{\alpha/2,r-k-1}$ e $\chi^2_{1-\alpha/2,r-k-1}$ sono le soluzioni di:

$$P(Q < \chi^2_{1-\alpha/2,r-k-1}) = \frac{\alpha}{2}, \qquad P(Q < \chi^2_{\alpha/2,r-k-1}) = 1 - \frac{\alpha}{2}$$



Applicazione su Poisson

- In un incrocio stradale sono stati registrati il numero di incidenti che si sono verificati ogni giorno per un totale di 75 giorni distinti

```
> camppois<-c(0, 3, 2, 0, 1, 2, 1, 1, 0, 1, 0, 1, 0, 0, 0,
+ 0, 0, 1, 0, 2, 0, 1, 0, 0, 0, 0, 1, 1, 3, 2,
+ 0, 1, 0, 1, 1, 0, 2, 3, 2, 1, 0, 0, 0, 1, 0,
+ 0, 0, 1, 0, 3, 0, 1, 0, 2, 4, 2, 0, 1, 1, 3,
+ 1, 0, 1, 0, 0, 0, 1, 0, 2, 4, 2, 0, 1, 2, 3)
>
> n<-length(camppois)
> n
[1] 75
>
> freq<-table(camppois)
> freq
camppois
 0  1  2  3  4
34 22 11  6  2
```

Applicazione su Poisson

- In un incrocio stradale sono stati registrati il numero di incidenti che si sono verificati ogni giorno per un totale di 75 giorni distinti

```
> camppois<-c(0, 3, 2, 0, 1, 2, 1, 1, 0, 1, 0, 1, 0, 0, 0,
+ 0, 0, 1, 0, 2, 0, 1, 0, 0, 0, 0, 1, 1, 3, 2,
+ 0, 1, 0, 1, 1, 0, 2, 3, 2, 1, 0, 0, 0, 1, 0,
+ 0, 0, 1, 0, 3, 0, 1, 0, 2, 4, 2, 0, 1, 1, 3,
+ 1, 0, 1, 0, 0, 0, 1, 0, 2, 4, 2, 0, 1, 2, 3)
>
> n<-length(camppois)
> n
[1] 75
>
> freq<-table(camppois)
> freq
camppois
 0  1  2  3  4
34 22 11  6  2
```

l'ampiezza del campione è n = 75 e corrisponde al numero di giorni considerati

Applicazione su Poisson

- In un incrocio stradale sono stati registrati il numero di incidenti che si sono verificati ogni giorno per un totale di 75 giorni distinti

```
> camppois<-c(0, 3, 2, 0, 1, 2, 1, 1, 0, 1, 0, 1, 0, 0, 0,
+ 0, 0, 1, 0, 2, 0, 1, 0, 0, 0, 0, 1, 1, 3, 2,
+ 0, 1, 0, 1, 1, 0, 2, 3, 2, 1, 0, 0, 0, 1, 0,
+ 0, 0, 1, 0, 3, 0, 1, 0, 2, 4, 2, 0, 1, 1, 3,
+ 1, 0, 1, 0, 0, 0, 1, 0, 2, 4, 2, 0, 1, 2, 3)
>
> n<-length(camppois)
> n
[1]  75
>
> freq<-table(camppois)
> freq
camppois
 0  1  2  3  4
34 22 11  6  2
```

l'ampiezza del campione è n = 75 e corrisponde al numero di giorni considerati

Nei 75 giorni nell'incrocio stradale in esame si sono verificati:

- 0 incidenti in 34 giorni
- 1 incidente in 22 giorni
- 2 incidenti in 11 giorni
- 3 incidenti in 6 giorni
- 4 incidenti in 2 giorni

Applicazione su Poisson

- Vogliamo verificare se il numero di incidenti sia descrivibile con una variabile aleatoria X di Poisson di parametro λ ($\lambda > 0$), ossia

$$p_X(x) = \frac{\lambda^x}{x!} e^{-\lambda} \qquad (x = 0, 1, \dots)$$

- I dati del campione permettono di ottenere una stima del parametro λ
  - Ricordando che uno stimatore corretto con varianza uniformemente minima del parametro λ di una distribuzione di Poisson risulta essere la **media campionaria**, si ha

```
> stimalambda<-mean(camppois)
> stimalambda
[1] 0.9333333
```

  - Supponiamo di considerare 4 categorie corrispondenti agli intervalli:

$$I_1 = \{0\}, \qquad I_2 = (0,1], \qquad I_3 = (1,2], \qquad I_4 = (2,+\infty)$$

- Le probabilità associate agli intervalli sono:

$$p_1 = p_X(0), \qquad p_2 = p_X(1), \qquad p_3 = p_X(2) \qquad p_4 = 1 - p_X(0) - p_X(1) - p_X(2)$$

Applicazione su Poisson

- Supponiamo di considerare 4 categorie corrispondenti agli intervalli:

$$I_1 = \{0\}, \qquad I_2 = (0,1], \qquad I_3 = (1,2], \qquad I_4 = (2,+\infty)$$

- Le probabilità associate agli intervalli sono:

$$p_1 = p_X(0), \qquad p_2 = p_X(1), \qquad p_3 = p_X(2) \qquad p_4 = 1 - p_X(0) - p_X(1) - p_X(2)$$

- Possiamo calcolarle:

```
> p<-numeric(4)
> p[1]<-dpois(0,stimalambda)
> p[2]<-dpois(1,stimalambda)
> p[3]<-dpois(2,stimalambda)
> p[4]<-1-p[1]-p[2]-p[3]
> p
[1] 0.39324072 0.36702467 0.17127818 0.06845643
```

- Il numero di elementi del campione nei quattro intervalli

```
> min(n*p[1],n*p[2],n*p[3],n*p[4])
[1] 5.134232
> r<-4
> nint<-numeric(r)
> nint[1]<-length(which(camppois==0))
> nint[2]<-length(which(camppois==1))
> nint[3]<-length(which(camppois==2))
> nint[4]<-length(which(camppois>2))
> nint
[1] 34 22 11  8
> sum(nint)
[1] 75
```

Applicazione su Poisson

- Calcoliamo χ^2:

```
> chi2<-sum(((nint-n*p)/sqrt(n*p))^2)
> chi2
[1] 3.663227
```

$$\sum_{i=1}^{r}\left(\frac{n_i - np_i}{\sqrt{np_i}}\right)^2$$

- Cioè $\chi^2 = 3{,}663227$
- Poiché:
  - il numero di categorie (o Intervalli) è $r = 4$
  - Ponendo $k = 1$ poichè la probabilità di Poisson contiene un parametro non noto si ha $r - k - 1 = 2$ e considerando $\alpha = 0.01$, calcoliamo

```
> r<-4
> k<-1
> alpha<-0.01
> qchisq(alpha/2,df=r-k-1)
[1] 0.01002508
> qchisq(1-alpha/2,df=r-k-1)
[1] 10.59663
```

Applicazione su Poisson

- Calcoliamo χ^2:

```
> chi2<-sum(((nint-n*p)/sqrt(n*p))^2)
> chi2
[1] 3.663227
```

- Cioè $\chi^2 = 3{,}663227$
- Poiché:
  - il numero di categorie (o Intervalli) è $r = 4$
  - Ponendo $k = 1$ poichè la probabilità di Poisson contiene un parametro non noto si ha $r - k - 1 = 2$ e considerando $\alpha = 0.01$, calcoliamo

```
> r<-4
> k<-1
> alpha<-0.01
> qchisq(alpha/2,df=r-k-1)
[1] 0.01002508
> qchisq(1-alpha/2,df=r-k-1)
[1] 10.59663
```

$\chi^2_{1-\alpha/2,r-k-1} = 0.010$ e $\chi^2_{\alpha/2,r-k-1} = 10.597$

- Essendo $0.010 < \chi^2 < 10.597$, l'ipotesi $\mathbf{H_0}$ di popolazione di Poisson **può essere accettata**

Applicazione sulla Normale

- Un urbanista è interessato alla superficie media μ delle abitazioni di una certa città
- A questo scopo osserva un campione di 50 appartamenti

```
> campnorm<-c(112.6, 118.2, 124.8, 122.1, 137.5, 106.7, 123.7,
+ 127.3, 123.2, 125.1, 120.8, 112.9, 117.0, 128.1, 102.9, 119.1,
+ 127.2, 124.8, 118.0, 131.4, 117.0, 118.2, 125.8, 116.2, 118.5,
+ 120.8, 127.1, 125.0, 131.2, 120.2, 126.0, 119.2, 112.4, 124.6,
+ 117.7, 116.1, 125.3, 115.5, 129.6, 119.1, 130.6, 125.3, 128.7,
+ 134.6, 124.5, 117.2, 126.1, 116.1, 116.0, 125.6)
> n<-length(campnorm)
> n
[1]  50
>
> m<-mean(campnorm)
> m
[1]  121.872
> d<-sd(campnorm)
> d
[1]  6.735469
```

Applicazione sulla Normale

- Un urbanista è interessato alla superficie media μ delle abitazioni di una certa città
- A questo scopo osserva un campione di 50 appartamenti

```
> campnorm<-c(112.6, 118.2, 124.8, 122.1, 137.5, 106.7, 123.7,
+ 127.3, 123.2, 125.1, 120.8, 112.9, 117.0, 128.1, 102.9, 119.1,
+ 127.2, 124.8, 118.0, 131.4, 117.0, 118.2, 125.8, 116.2, 118.5,
+ 120.8, 127.1, 125.0, 131.2, 120.2, 126.0, 119.2, 112.4, 124.6,
+ 117.7, 116.1, 125.3, 115.5, 129.6, 119.1, 130.6, 125.3, 128.7,
+ 134.6, 124.5, 117.2, 126.1, 116.1, 116.0, 125.6)
> n<-length(campnorm)
> n
[1] 50
>
> m<-mean(campnorm)
> m
[1] 121.872
> d<-sd(campnorm)
> d
[1] 6.735469
```

→ Media Campionaria

→ Deviazione Standard Campionaria

Applicazione sulla Normale

- Vogliamo verificare se la popolazione da cui proviene il campione può essere descritta da una variabile aleatoria X di densità normale
  - Applichiamo il test chi-quadrato di misura $\alpha = 0.05$

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}, \qquad x \in \mathbb{R} \qquad (\mu \in \mathbb{R}, \sigma > 0)$$

- Supponiamo di suddividere l'insieme dei valori che X può assumere in $r = 5$ sottoinsiemi $I_1, I_2, \dots, I_5$ in modo che risulti
  - Che la probabilità che X assuma un valore appartenente a I_i $(i = 1, 2, \dots, 5)$ sia uguale a $p_i = 0.2$
- Ricordando che:
  - Lo stimatore di μ è la media campionaria
  - Lo stimatore di σ^2 è la varianza campionaria
    - Utilizzando i quantili della normale possiamo determinare i sottoinsiemi $I_1, I_2, \dots, I_5$

Applicazione sulla Normale

- Utilizzando i quantili della normale possiamo determinare i sottoinsiemi $I_1, I_2, \dots, I_5$

```
> a<-numeric(4)
> for(i in 1:4)
+ a[i]<-qnorm(0.2*i,mean=m,sd=d)
> a
[1] 116.2033 120.1656 123.5784 127.5407
```

- Si ha che gli intervalli sono:

$$I_1 = (-\infty, 116.20), \quad I_2 = [116.2, 120.17), \quad I_3 = [120.17, 123.58),$$
$$I_4 = [123.58, 127.54), \quad I_5 = [127.54, +\infty).$$

- Determinare il numero di elementi del campione che cadono nei singoli intervalli:

```
> r<-5
> nint<-numeric(r)
> nint[1]<-length(which(campnorm<a[1]))
> nint[2]<-length(which((campnorm>=a[1])&(campnorm<a[2])))
> nint[3]<-length(which((campnorm>=a[2])&(campnorm<a[3])))
> nint[4]<-length(which((campnorm>=a[3])&(campnorm<a[4])))
> nint[5]<-length(which(campnorm>=a[4]))
> nint
[1] 10 11  5 16  8
> sum(nint)
[1] 50
```

Le frequenze degli intervalli sono:
$n_1 = 10 \;\; n_2 = 11 \;\; n_3 = 5 \;\; n_4 = 16 \;\; n_5 = 8$

Applicazione sulla Normale

- Calcoliamo χ^2:

```
> chi2<-sum(((nint-n*0.2)/sqrt(n*0.2))^2)
> chi2
[1] 6.6
```

$$\sum_{i=1}^{r}\left(\frac{n_i - np_i}{\sqrt{np_i}}\right)^2$$

- Cioè $\chi^2 = 6{,}6$
- Poiché:
  - il numero di categorie (intervalli) è $r = 5$
  - Ponendo $k = 3$ poiché la probabilità Normale contiene due parametri non noti

    si ha $r - k - 1 = 2$ e considerando $\alpha = 0.05$, calcoliamo

```
> r<-5
> k<-2
> alpha<-0.05
> qchisq(alpha/2,df=r-k-1)
[1] 0.05063562
> qchisq(1-alpha/2,df=r-k-1)
[1] 7.377759
```

$\chi^2_{1-\alpha/2,r-k-1} = 0.0506$ e $\chi^2_{\alpha/2,r-k-1} = 7.378$

Essendo $0.0506 < \chi^2 < 7.378$, l'ipotesi $\mathbf{H_0}$ di popolazione normale può essere accettata