Corso di Laurea Magistrale in Informatica

Statistica e analisi dei dati

Inferenza statistica

Professoressa:

Amelia G. Nobile

Studentessa:

Anna Tomeo

Anno 2018/2019

Inferenza statistica

L'indagine statistica è sempre effettuata su un insieme detto popolazione. La conoscenza delle caratteristiche di una popolazione può essere ottenuta osservando la totalità delle entità della popolazione oppure un sottoinsieme di questa, detto campione estratto dalla popolazione. Per studiare l'inferenza statistica occorre conoscere le varie variabili aleatorie X discrete e continue.

Di particolare importanza in statistica è **l'inferenza statistica**. Essa ha lo scopo di estendere le misure ricavate dall'esame di un campione alla popolazione da cui il campione è stato estratto.

Uno dei problemi centrali dell'inferenza statistica è il seguente: si desidera studiare una popolazione descritta da una variabile aleatoria osservabile X la cui funzione di distribuzione ha una forma nota ma contiene un parametro $\vartheta \in \Theta$ non noto (o più parametri non noti).

osservabile: si possono ricavare i valori assunti dalla variabile aleatoria X tramite, ad esempio, l'esecuzione di un esperimento casuale. Ovviamente se ϑ è noto la legge di probabilità è completamente specificata.

Per ottenere informazioni sul parametro non noto di una popolazione si fa uso dell'inferenza statistica, considerando un campione rappresentativo della popolazione.

L'inferenza statistica si basa su due metodi fondamentali di indagine: **la stima dei parametri** e la **verifica delle ipotesi.**

La stima dei parametri ha lo scopo di determinare i valori non noti dei parametri di una popolazione (come il valore medio, la varianza,...) per mezzo dei corrispondenti parametri derivati dal campione estratto dalla popolazione (come la media campionaria, la varianza campionaria,...). Si possono usare stime puntuali o stime per intervallo.

Si parla di stima puntuale quando si stima un parametro non noto di una popolazione usando un singolo valore reale.

Con le stime intervallari (o intervallo di confidenza) si cerca di determinare in base al campione osservato (x_1,..., x_n) due limiti (uno inferiore e uno superiore) entro i quali sia compreso il parametro non noto con un certo grado di confidenza, detto anche grado di fiducia.

La verifica delle ipotesi è un procedimento che consiste nel fare una congettura o un'ipotesi sul parametro non noto ϑ e decidendo sulla base del campione estratto se essa è accettabile.

Per affrontare questi problemi occorrono le variabili aleatorie discrete o continue.

Una variabile aleatoria è discreta se l'insieme dei valori è finito o numerabile.

Una variabile aleatoria è continua se l'insieme dei valori è più numeroso o per meglio dire in un determinato intervallo.

Nella relazione da me presentata si è deciso di approfondire una variabile aleatoria continua, in particolare la **variabile aleatoria normale**.

Variabile aleatoria normale

La funzione di distribuzione normale, detta anche di Gauss o gaussiana, riveste estrema importanza nel calcolo delle probabilità e nella statistica anche in quanto essa costituisce una distribuzione limite alla quale tendono varie altre funzioni di distribuzioni sotto opportune ipotesi.

Una variabile aleatoria X di densità di probabilità

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

Si dice avere distribuzione normale di parametri μ e σ.

Prendiamo in esame in seguente vettore di dati che contiene i voti di 100 studenti.

```
> voti<-sample(c(18:30), size=100, replace=TRUE)
```

```
 [1] 28 30 28 23 20 28 25 27 28 21 23 23 19 26 19 19 19 27 19 23 20 21 29 19 27
[26] 19 20 22 19 20 29 30 28 18 23 24 20 27 22 28 28 29 19 18 28 30 23 20 18 26
[51] 24 28 24 23 22 29 25 27 22 20 18 27 19 29 23 30 18 25 26 18 28 25 28 18 18
[76] 18 28 23 18 25 21 26 27 30 27 26 22 29 22 20 27 25 26 28 22 29 18 28 30 19
```

Calcoliamo la densità normale di una variabile aleatoria normale:

```
> media<-mean(voti)
```

```
> media
```

```
[1] 23.87
```

```
> ds<-sd(voti)
```

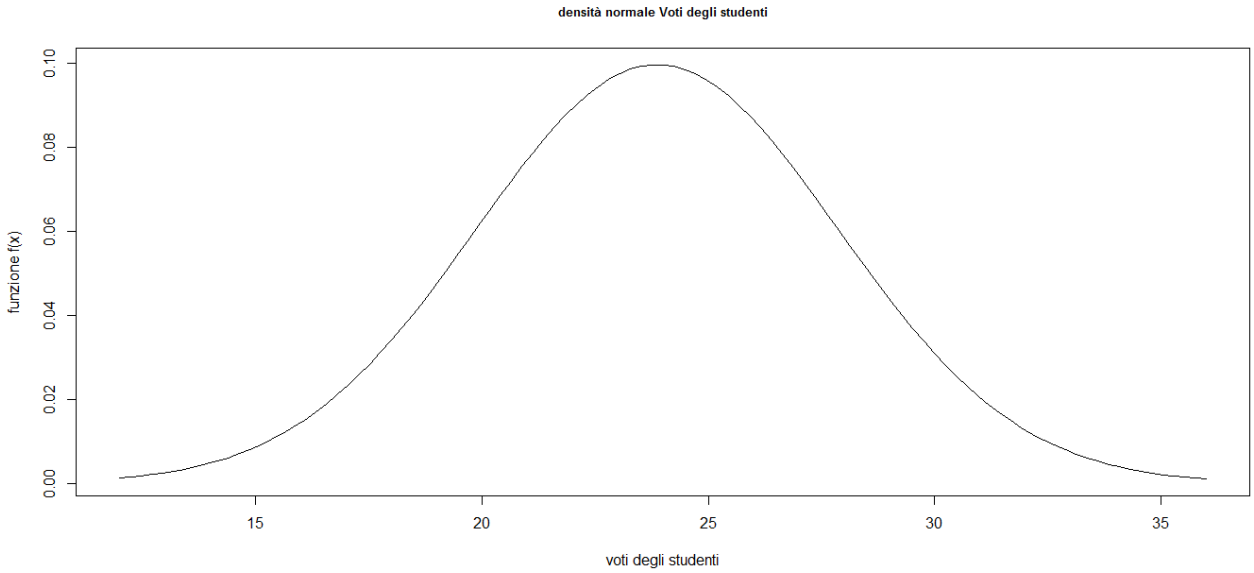
```
> ds
```

```
[1] 4.004177
```

```
> x<-voti
```

In R la densità normale si calcola con dnorm.

```
> curve (dnorm (x, media, ds), from=12, to=36, xlab="voti degli studenti", ylab="funzione f(x)", main="densità normale Voti degli studenti", cex.main=0.8)
```



La funzione curve() ha al suo interno i seguenti parametri: dnorm() avente come parametri in ingresso il valore medio e la deviazione standard; from e to che indicano il range entro il quale disegnare il grafico. L'intervallo to-from è stato calcolato utilizzando *la* ***regola del 3σ****,* in *maniera tale che l'area sottesa dalla curva sia circa unitaria e l'area delle code destra e sinistra sia trascurabile.*

$$P(\mu-3\,\sigma < X < \mu+3\,\sigma) = P\left(-3 < \frac{X-\mu}{\sigma} < 3\right) = P(-3 < Z < 3) = 0.9973002.$$

Quindi, la probabilità che la variabile assuma valori in un intervallo avente come centro μ e semiampiezza 3σ è prossima all'unità. Andiamo a verificare nel nostro caso tramite il comando pnorm():

> pnorm(36, media, ds)-pnorm(12, media, ds)

[1] 0.9972582

Tale valore è prossimo all'unità e perciò possiamo dire che il nostro intervallo è corretto.

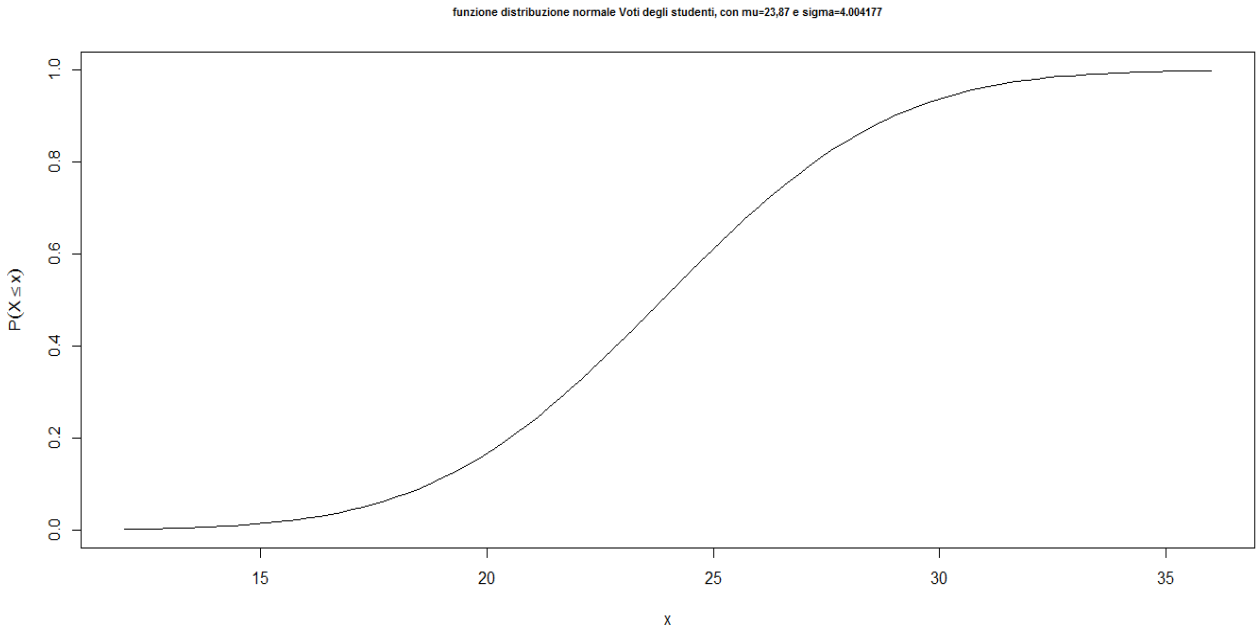
Il punto più alto della curva si trova proprio in corrispondenza del valore medio, ossia 23, 87. La deviazione standard è 4, infatti la curva non è molto larga. **Se la deviazione standard fosse stata bassa la curva sarebbe stata più stretta**.

Calcoliamo la funzione di distribuzione normale che è definita come:

$$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f_X(y)\,dy = \Phi\left(\frac{x-\mu}{\sigma}\right) \qquad x \in \mathbb{R}$$

In R si calcola con pnorm():

> curve(pnorm(x, media, ds), from=12, to=36, xlab="x", ylab=expression (P(X<=x)), main="funzione distribuzione normale Voti degli studenti, con mu=23,87 e sigma=4.004177", cex.main=0.65)



Notiamo che per i voti che vanno da 20 a 26 circa la probabilità aumenta (la curva tende ad alzarsi).

In R è possibile calcolare i quantili per una distribuzione normale, tramite qnorm:

> z<-c(0, 0.25, 0.5, 0.75, 1)

> qnorm(z, media, ds)

[1] -Inf 21.16922 23.87000 26.57078 Inf

Il voto 21 corrisponde al 25% dei voti, il voto 27 corrisponde al 75% e la media (ossia il 50% dei voti) corrisponde proprio a 23.87. Q0= -inf e Q4= +inf.

Stima puntuale

Dopo aver osservato i valori assunti dalla popolazione si è interessati a stimare i parametri non noti della stessa.

Ci sono due metodi per poter effettuare tali stime:

1) *metodo dei momenti;*
2) *metodo della massima verosomiglianza.*

Scegliamo di trattare il metodo dei momenti.

Il metodo dei momenti quindi fornisce:

• Uno stimatore per il valore medio μ che corrisponde alla media campionaria (X_1+ X_2+...+ X_n)/n

• Uno stimatore per la varianza σ^2 =(n-1) * S^2/n

In R:

> stimaMu<-mean(voti)

> stimaMu

[1] 23.87

> sigmaQ<-(length(voti)-1)*var(voti)/length(voti)

> sigmaQ

[1] 15.8731

Vediamo che lo stimatore per il valore medio μ = 23.87 e lo stimatore della varianza σ^2=15.8731.

(X_1+ X_2+...+ X_n)/n è uno stimatore corretto con varianza uniformemente minima

Per una popolazione normale lo stimatore (n-1) * S^2/n della varianza individuato con il metodo dei momenti è asintoticamente corretto.

Stime intervallari

Alla stima puntuale di un parametro non noto si preferisce sostituire un intervallo di valori entro il quale sia compreso il parametro non noto della popolazione con un certo coefficiente di confidenza o detto anche grado di fiducia.

Sia $X_1,\ldots,X_n$ un campione casuale di ampiezza n estratto da una popolazione (discreta o continua) e sia ϑ il parametro non noto. Denotiamo con $\underline{C}_n = g_1(X_1,\ldots,X_n)$ e con $\overline{C}_n = g_2(X_1,\ldots,X_n)$ due statistiche (funzioni osservabili del campione casuale) che soddisfino la condizione $\underline{C}_n < \overline{C}_n$.

Fissato un coefficiente di confidenza $1-\alpha$ ($0 < \alpha < 1$), è possibile scegliere le statistiche $\underline{C}_n$, $\overline{C}_n$, in modo tale che

$$P(\underline{C}_n < \vartheta < \overline{C}_n) = 1 - \alpha,$$

allora si dice che $(\underline{C}_n, \overline{C}_n)$ è un intervallo di confidenza di grado $1-\alpha$ per ϑ. Inoltre, le statistiche $\underline{C}_n$, $\overline{C}_n$ sono dette limite inferiore e superiore dell'intervallo di confidenza.

Per quanto riguarda la distribuzione normale esistono vari casi di stima dei parametri non noti:

1) determinare un intervallo di confidenza di grado 1- α per il valore medio μ nel caso in cui la varianza della popolazione normale è nota;
2) determinare un intervallo di confidenza di grado 1- α per il valore medio μ nel caso in cui la varianza della popolazione normale è non nota;
3) determinare un intervallo di confidenza di grado 1- α per la varianza nel caso in cui il valore medio μ della popolazione normale è noto;
4) determinare un intervallo di confidenza di grado 1- α per la varianza nel caso in cui il valore medio della popolazione normale è non noto;

Osservazione: i casi 2 e 4 sono realistici, mentre 1 e 3 sono teorici.

Per la distribuzione normale avremo svariati casi tutti riconducibili a un grafico del seguente tipo:

In questa relazione determiniamo un intervallo di confidenza di grado 1- α per il valore medio μ nel caso in cui **la varianza della popolazione normale è non nota**;

Per risolvere tale problema usufruiremo della densità di Student con n-1 gradi di libertà:

$$T_n = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \sqrt{\frac{\sigma^2}{S_n^2}} = \frac{Z_n}{\sqrt{Q_n/(n-1)}},$$

Definiamo brevemente il metodo pivotale.

Il metodo pivotale è un metodo per la costruzione di intervalli di confidenza. Tale metodo consiste essenzialmente nel determinare una variabile aleatoria di pivot $\gamma(X_1,\ldots,X_n;\vartheta)$ che dipende dal campione casuale $X_1,\ldots,X_n$ e dal parametro non noto ϑ e la cui funzione di distribuzione non contiene il parametro da stimare. Tale variabile aleatoria non è una statistica poichè dipende dal parametro non noto ϑ e quindi non è osservabile. Per determinare un intervallo di confidenza di grado 1- α per il problema 2 utilizziamo il metodo pivotale con la seguente variabile aleatoria di pivot.

$$T_n = \frac{\overline{X}_n - \mu}{S_n/\sqrt{n}}.$$

Tale variabile aleatoria dipende dal campione casuale e dal parametro non noto e, quindi, può essere interpretata come una variabile aleatoria di pivot. Inoltre, $T_n$abbiamo visto che è distribuita con legge di Student con n-1 gradi di libertà.

Applicando il metodo pivotale per tale variabile aleatoria siamo quindi in grado di ricavare l'intervallo:

$$P\left(T_n < -t_{\alpha/2,n-1}\right) = P\left(T_n > t_{\alpha/2,n-1}\right) = \frac{\alpha}{2},$$

$$P\left(-t_{\alpha/2,n-1} < T_n < t_{\alpha/2,n-1}\right) = 1 - \alpha.$$

Dove i limiti inferiori e superiori corrispondono a $\frac{\alpha}{2}$, mentre la regione di accettazione corrisponde a 1-α. Il grado di fiducia 1-α attribuito alla stima è scelto da un DECISORE.

Siccome questa volta abbiamo a che fare con una variabile di Student, il comando per calcolare i limiti inferiori e superiori è qt().

> alpha<-1-0.99

> n<-length(voti)

> qt(1-alpha/2, df=n-1) #calcoliamo il limite superiore, al quale aggiungiamo segno negativo per il limite inferiore

[1] 2.626405

> mean(voti)-qt(1-alpha/2, df=n-1)*sd(voti)/sqrt(n)

[1] 22.81834

> mean(voti)+qt(1-alpha/2, df=n-1)*sd(voti)/sqrt(n)

[1] 24.92166

Assegnato il grado di accettazione abbiamo bisogno di calcolare i quantili di una popolazione normale: il calcolo dei quantili (comando qt() con n-1 gradi libertà dove n indica la lunghezza del campione studiato) delinea nel nostro grafico quali sono i rispettivi limiti inferiori e superiori della regione di accettazione.

Il limite inferiore della regione di accettazione = -2.626405 e il limite superiore = 2.626405.

Il nostro intervallo di confidenza è (22.81834, 24.92166). Il nostro valore medio = 23.87, quindi è compreso nella stima dell'intervallo, ossia accettato.

Differenza tra valori medi di una popolazione normale

Vogliamo ora costruire degli intervalli di confidenza per la differenza tra i valori medi di due popolazioni normali.

Per determinare un intervallo di confidenza di grado $1 - \alpha$ per $\mu_1 - \mu_2$ quando entrambe le varianze σ_1^2 e σ_2^2 delle due popolazioni normali sono note, consideriamo la variabile aleatoria di pivot:

$$Z_n = \frac{\overline{X}_{n_1} - \overline{Y}_{n_2} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}.$$

dove, X_{n1}(segnato) e X_{n2}(segnato) sono le medie campionarie delle osservazioni.

Tale variabile aleatoria dipende dal campione casuale e dal parametro non noto $\mu_1 - \mu_2$ (le varianze campionarie σ_1^2 e σ_2^2 delle due popolazioni sono note) ed è caratterizzata da una densità normale standard.

Applicando il metodo pivotale diremo che la nostra regione di accettazione sarà del tipo:

$$P\left(-z_{\alpha/2} < \frac{\overline{X}_{n_1} - \overline{Y}_{n_2} - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} < z_{\alpha/2}\right) = 1 - \alpha,$$

Che in questo caso si riduce a:

$$\overline{x}_{n_1} - \overline{y}_{n_2} - z_{\alpha/2}\sqrt{\frac{s_{n_1}^2}{n_1} + \frac{\tilde{s}_{n_2}^2}{n_2}} < \mu_1 - \mu_2 < \overline{x}_{n_1} - \overline{y}_{n_2} + z_{\alpha/2}\sqrt{\frac{s_{n_1}^2}{n_1} + \frac{\tilde{s}_{n_2}^2}{n_2}}.$$

Caso di studio:

determinare un dato intervallo di confidenza per $\mu_1 - \mu_2$ (che corrispondono al valore medio della prima e seconda popolazione) quando entrambe le varianze σ_1^2 e σ_2^2 sono note.

Abbiamo due insiemi di voti di studenti.

per A: campione di 50 voti, media=26.98 e deviazione standard=5;

per B: campione di 100 voti, media=23.87, deviazione standard=4.

> alpha<-1-0.99

> qnorm(1-alpha/2, mean=0, sd=1) #è una variabile normale standard quindi ha media=0 e deviazione standard=1

[1] 2.575829 = $z_{\frac{\alpha}{2}}$

> nA<-50

> nB<-100

> mA<-26.98

> mB<-23.7

> sigmaA<-5

> sigmaB<-4

> (mA-mB-qnorm(1-alpha, mean=0, sd=1)*sqrt(sigmaA^2/nA+sigmaB^2/nB))

1.390066

> (mA-mB+qnorm(1-alpha, mean=0, sd=1)*sqrt(sigmaA^2/nA+sigmaB^2/nB))

5.169934

La stima per l'intervallo di confidenza di grado 1- α=0.99 per le differenze tra i valori medi μ_1 -μ_2 è (1.390066, 5.169934). Sia il limite inferiore che quello superiore sono positivi, il che significa che il primo gruppo è migliore del secondo.

Verifica delle ipotesi

Le aree più importanti dell'inferenza statistica sono la stima dei parametri e la verifica delle ipotesi. La verifica delle ipotesi interviene spesso nelle ricerche di mercato, nelle indagini sperimentali e industriali, nei sondaggi di opinione, ecc. In generale gli elementi che costituiscono il punto di partenza del procedimento di verifica delle ipotesi sono una popolazione descritta da una variabile aleatoria X caratterizzata da una funzione di probabilità o densità di probabilità f(x; ϑ), un'ipotesi su di un parametro non noto della popolazione ed un campione casuale x_1, x_2,..., x_n, estratto dalla popolazione. Occorre in primo luogo precisare il significato di ipotesi statistica.

Un'ipotesi statistica è un'affermazione o una congettura sul parametro non noto ϑ. Se l'ipotesi statistica specifica completamente f(x;ϑ) è detta ipotesi semplice, altrimenti è chiamata ipotesi composta.

Per denotare un'ipotesi statistica useremo il carattere H seguito dai due punti e successivamente dall'affermazione che specifica l'ipotesi.

I test statistici sono di due tipi: test unilaterali (detti anche unidirezionali) e test bilaterali (detti anche bidirezionali). Un test bilaterale è il seguente:

$H_0 : \mu = \mu_0$

$H_1 : \mu \neq \mu_0$

mentre test unilaterali sono i seguenti:

(sinistro)

$H_0 : \mu \leq \mu_0$

$H_1 : \mu > \mu_0$

(destro)

$H_0 : \mu \geq \mu_0$

$H_1 : \mu < \mu_0$

Problemi esistenti:

1) Verifica di ipotesi sul valore medio μ nel caso in cui la varianza σ^2 della popolazione normale è nota;
2) Verifica di ipotesi sul valore medio μ nel caso in cui la varianza della popolazione normale è non nota;
3) Verifica di ipotesi sulla varianza σ^2 nel caso in cui il valore medio μ della popolazione normale è noto;

4) Verifica di ipotesi sulla varianza σ^2 nel caso in cui il valore medio della popolazione normale è non noto.

Il nostro caso di studio riguarda il secondo problema:

Test su μ con varianza non nota

Test bilaterale: Sia (x_1,..., x_n,) un campione osservato di ampiezza n estratto da una popolazione normale con varianza non nota. Si considerino le ipotesi:

H_0 : μ = μ_0

H_1 : μ ≠ μ_0

Essendo la varianza non nota, entrambe le ipotesi sono composte. Quando H_0 è vera, in analogia a quanto visto per gli intervalli di confidenza, gioca un ruolo fondamentale la variabile aleatoria

$$T_n = \frac{\overline{X}_n - \mu_0}{S_n/\sqrt{n}}.$$

che è distribuita con legge di Student con n−1 gradi di libertà.

$$\text{- si accetti } \mathbf{H_0} \text{ se} \quad -t_{\alpha/2,n-1} < \frac{\overline{x}_n - \mu_0}{s_n/\sqrt{n}} < t_{\alpha/2,n-1}$$

$$\text{- si rifiuti } \mathbf{H_0} \text{ se} \quad \frac{\overline{x}_n - \mu_0}{s_n/\sqrt{n}} < -t_{\alpha/2,n-1} \quad \text{oppure} \quad \frac{\overline{x}_n - \mu_0}{s_n/\sqrt{n}} > t_{\alpha/2,n-1}$$

Il nostro caso:

Una scuola sostiene che la media dei voti dei suoi studenti sia 24. Si prende un campione di 100 studenti e si verifica che hanno una media di 23.87.

Si desidera utilizzare il test di misura α = 0.01 per verificare l'ipotesi nulla H_0 : μ = 24 in alternativa all'ipotesi H_1: μ ≠ 24.

Utilizziamo il test d verifica bilaterale utilizzando R:

> alpha<-0.01

> mu0<-24

> n=100

> qt(1-alpha/2, df=n-1)

2.626405

> meancamp<-23.87

> devcamp<-4

> (meancamp-mu0)/(devcamp/sqrt(n))

> -0.325

$t_{\frac{\alpha}{2},n-1}$ = 2.626405

Quindi -0.325 è nella regione di accettazione. Quindi occorre accettare l'ipotesi nulla.

Test del chi-quadrato

In molti problemi reali, si desidera 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). A questo scopo, utilizzeremo il criterio di verifica delle ipotesi del chi–quadrato, detto anche test del chi–quadrato o test del buon adattamento.

Il test del chi-quadrato mira a rispondere a due ipotesi:

H_0: X ha funzione di distribuzione F_x(x) (avendo stimato k parametri non noti in base al campione);

H_1: X non ha una funzione di distribuzione F_x(x).

Il test di verifica delle ipotesi considerato è bilaterale.

Come si fa:

Suddividiamo l'insieme dei valori che la variabile aleatoria X può assumere in r sottoinsiemi I_1,..., I_r in modo che risulti essere uguale a pi la probabilità che, secondo la distribuzione ipotizzata, la variabile aleatoria assuma un valore appartenente a I_i.

Si estrae poi un campione x_1,..., $x_n$di ampiezza n e si osservano le frequenze assolute n_1,..., n_r con cui gli n elementi si distribuiscono nei rispettivi insiemi I_1,..., I_r. Quindi n_i rappresenta il numero degli elementi del campione che cadono nell'intervallo I_i (i = 1,2,...,r).

$$p_i \geq 0 \quad (i = 1, 2, \ldots, r), \qquad \sum_{i=1}^{r} p_i = 1;$$

$$n_i \geq 0 \quad (i = 1, 2, \ldots, r), \qquad \sum_{i=1}^{r} n_i = n.$$

La probabilità che esattamente n_1 elementi appartengano ad I_1, n_2 elementi appartengano ad I_2, ecc è una funzione di probabilità multinomiale:

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

Ne segue che il numero medio di elementi che cadono nell'intervallo I_i è np_i. Si calcola poi la quantità

$$\chi^2 = \sum_{i=1}^{r} \left(\frac{n_i - n\, p_i}{\sqrt{n\, p_i}} \right)^2.$$

Il criterio del chi-quadrato si basa sulla statistica

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

dove N_i è la variabile aleatoria che descrive il numero degli elementi del campione casuale X_1,..., X_n che cadono nell'intervallo I_i. (i=1,2,...r)

Se la variabile aleatoria X ha una funzione di distribuzione F_x(x) con k parametri non noti, si può dimostrare che per n sufficientemente grande la funzione di distribuzione della statistica Q è approssimabile con la funzione di distribuzione del chi-quadrato con r-k-1 gradi di libertà.

Per garantire che ogni classe contenga in media almeno 5 elementi, si ritiene valida l'approssimazione se risulta

$$\min(np_1, np_2, \ldots, np_r) \geq 5.$$

Caso di studio:

consideriamo il campione di voti che chiamiamo campnorm:

```
> m<-mean(campnorm)
> m
[1] 23.87
> d<-sd(campnorm)
> d
[1] 4.004177
> a<-numeric(4)
> for( i in 1:4)
+ a[i]<-qnorm(0.2*i, mean=m,sd=d)
```

Abbiamo utilizzato la funzione quantili della distribuzione normale: qnorm, dopo aver stimato mu e sigma tramite la media campionaria e la deviazione standard campionaria, per determinare i sottoinsiemi.

```
> a
```

[1] 20.50000 22.85555 24.88445 27.24000

Dai valori uscenti si possono ricavare gli intervalli dei 5 sottoinsiemi:

1. (-infinito, 20.50000)
2. (20.50000, 22.85555)
3. (22.85555, 24.88445)
4. (24.88445, 27.24000)
5. (27.24000, +infinito)

```
> 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] 30 10 12 21 27
> chi2<-sum(((nint -n*0.2)/sqrt(n*0.2))^2)
> chi2
[1] 15.7
> 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
```

Abbiamo calcolato la **lunghezza (n) del campione**, la **media campionaria** e la **deviazione standard campionaria**.

Abbiamo deciso di suddividere l'insieme dei valori che può assumere la variabile aleatoria in 5 sottoinsiemi (r=5), in modo che *un valore risulti avere una probabilità pari a 0.2* appartiene a un determinato sottoinsieme.

A questo punto bisogna calcolare il numero di elementi che cadono in un dato intervallo. Dopo alcuni calcoli abbiamo la seguente suddivisione:

in 1 cadono 30 elementi

in 2 cadono 10 elementi

in 3 cadono 12 elementi

in 4 cadono 21 elementi

in 5 cadono 27 elementi

il valore del chi-quadrato è pari a **15.7, ciò significa che va bene in quanto 15.7 ≥ 5.**

A questo punto bisogna calcolare i gradi di libertà cui deve essere applicata la densità chi-quadrato:

r-k-1 = 5-2-1=2

r: 5 sottoinsiemi;

k: numero di parametri non noti (mu e sigma).

Quindi abbiamo che la funzione di distribuzione del chi-quadrato **ha 2 gradi di libertà**.

I **limiti inferiori e superiore** sono rispettivamente **0.05063562 e 7.377759.**

Il valore del chi-quadrato 15.7 **non è compreso** in tale intervallo quindi per questo campione la distribuzione normale non può essere accettata.