Laboratorio di Probabilità e Statistica

lezione 6

Massimo Guerriero – Ettore Benedetti

Indice Lezione

- Prerequisiti dalla lezione scorsa
- Intervallo di confidenza per la media
- Verifica d'ipotesi sulla media
- Confronto tra le medie di gruppi
- Verifica di ipotesi di indipendenza

Prerequisiti dalla lezione scorsa

- Media e varianza campionaria
  - Legge dei grandi numeri
  - Teorema del limite centrale
- Intervallo di confidenza per la media

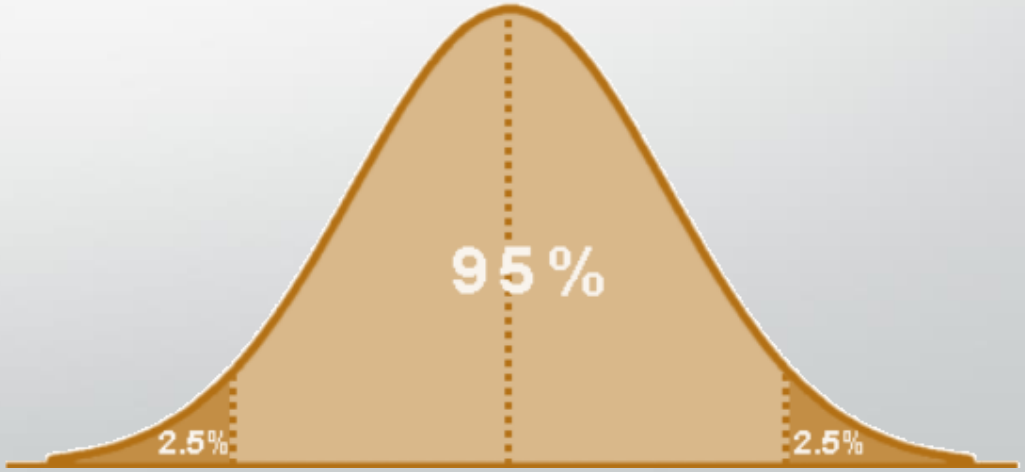
Intervallo di confidenza per la media 1/3

Gli intervalli di confidenza per la media forniscono un campo di variazione centrato sulla media campionaria all'interno del quale ci si aspetta di trovare il parametro incognito μ.

Per la variabile casuale normale standard $Z=\frac{\overline{X_n}-\mu}{\frac{\sigma}{\sqrt{n}}}$ con probabilità 1-α:

$$P\left(-z_{1-\frac{\alpha}{2}} < \frac{\overline{X_n}-\mu}{\frac{\sigma}{\sqrt{n}}} < z_{1-\frac{\alpha}{2}}\right) = 1-\alpha$$

$$P\left(\overline{X_n} - z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} < \mu < \overline{X_n} + z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right)$$



Intervallo di confidenza per la media 2/3

Nella maggior parte dei casi la varianza della popolazione non è conosciuta, quindi la si deve stimare. In questo caso la media campionaria si distribuisce come una v.a. t di Student con n-1 d.f.

$$\mu \in \left(\overline{X_n} \pm t^{(n-1)}_{1-\frac{\alpha}{2}} \sqrt{\frac{\overline{S_n^2}}{n}} \right)$$

In R si possono calcolare gli intervalli di confidenza manualmente (creandosi una funzione), oppure utilizzando il comando ***t.test****(campione, conf.lev = α)*

```
> x<-(c(0.39,0.68,0.82,1.35,1.38,1.62,
1.70,1.71,1.85,2.14,2.89,3.69))
> t.test(x, conf.lev=0.99)

 One Sample t-test

data:  x
t = 6.3305, df = 11, p-value = 5.595e-05
alternative hypothesis: true mean is not equal to 0
99 percent confidence interval:
 0.8583201 2.5116799
sample estimates:
mean of x
    1.685
```

Intervallo di confidenza per la media 3/3

Es. *Creare una funzione per calcolare gli intervalli di confidenza per la normale standard, accettando come parametri il campione, la varianza della popolazione da cui il campione è stato tratto, ed alfa.*

```
campione<-c(7.36, 11.91, 12.91, 9.77, 5.99, 10.91, 9.57, 11.01, 6.11, 12.12)

normalCI<-function(campione, varianza, alfa){
  xn<-mean(campione);
  sigma<-sqrt(varianza);
  z<-qnorm(1-(alfa/2));
  inf<-xn-(z*(sigma/sqrt(length(campione))));
  inf<-round(as.numeric(inf),3);
  sup<-xn+(z*(sigma/sqrt(length(campione))));
  sup<-round(as.numeric(sup),3);
  cat(paste("[",inf,",",sup,"]"));
}
```

Vediamo ora qualche dato sugli intervalli di confidenza presi da una distribuzione t di Student (varianza della popolazione incognita).

Tabella Excell

Consegna

1. Utilizzare la funzione NormalCI per verificare come cambiano le ampiezze degli intervalli di confidenza al variare di alfa (0.1, 0.05, 0.01) in un campione di 100 unità prese casualmente da una v.a. normale di media 20 e sd 3.

2. Verificare cosa succede se, a parità di alfa, numerosità del campione e media della popolazione, varia la sd della popolazione.

Indice Lezione

- Prerequisiti dalla lezione scorsa
- Intervallo di confidenza per la media
- Verifica d'ipotesi sulla media
- Confronto tra le medie di gruppi
- Verifica di ipotesi di indipendenza

Verifica d'ipotesi sulla media 1/6

Vogliamo rispondere a questa domanda:

- *disponendo di un campione di numerosità limitata, si può affermare che la media μ_x della popolazione da cui esso è stato estratto è diversa dal valore prestabilito μ_p?*

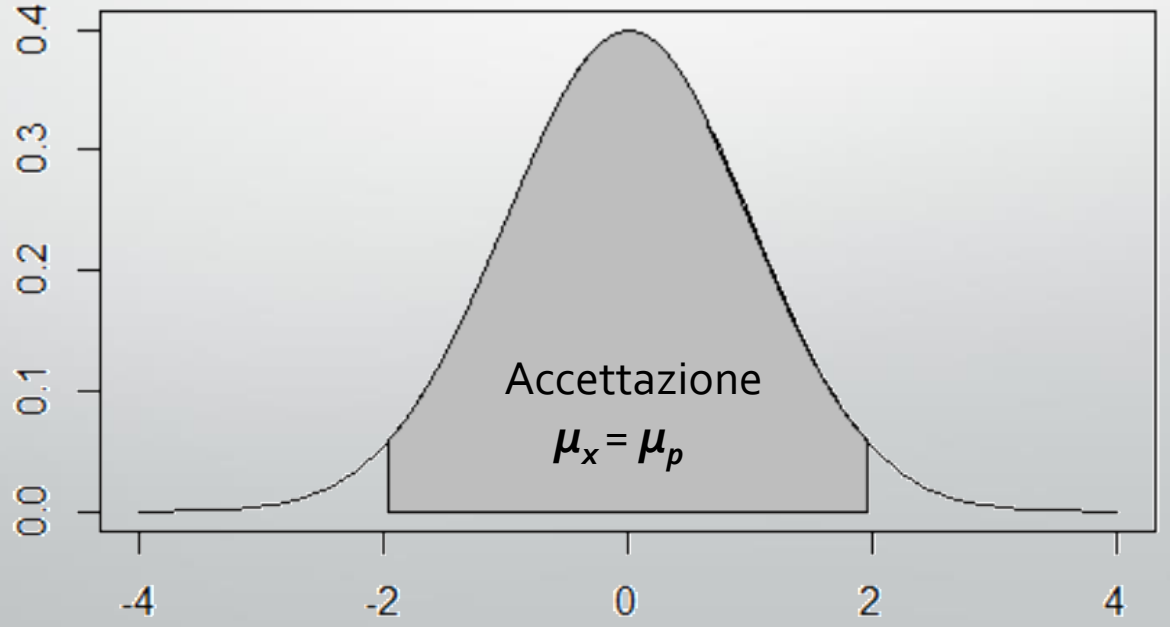
Per condurre il test si devono effettuare i seguenti tre passi:

1. Si fissa il "**tasso accettabile di rischio**" α. *Es. $\alpha = 0,05$.*
2. Si estrae il campione dalla popolazione e si determina la **media campionaria**
3. Si individua l'**intervallo di confidenza** ad 1- α mediante la variabile z (normale standard) se la varianza della popolazione è nota oppure mediante la t di Student se la varianza della popolazione è incognita.

Verifica d'ipotesi sulla media 2/6

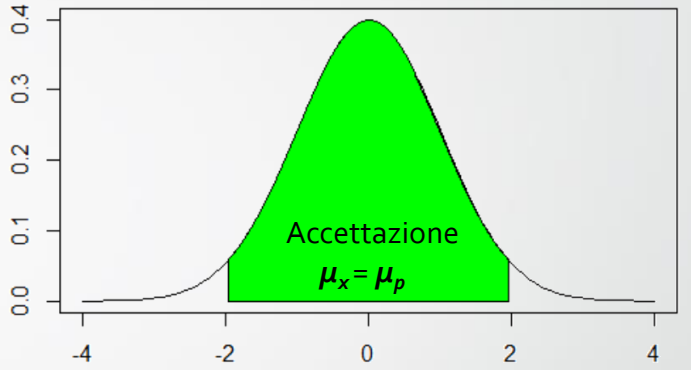
- Se z (oppure t) è compreso nell'intervallo di confidenza trovato NON si può affermare che μ_x sia diverso dal valore prestabilito μ_p .
- Se invece z (oppure t) NON è compreso nell'intervallo di confidenza trovato allora SI PUO' affermare, con una probabilità di errare non superiore ad α, che μ_x sia diverso dal valore prestabilito μ_p .

Codice Grafico

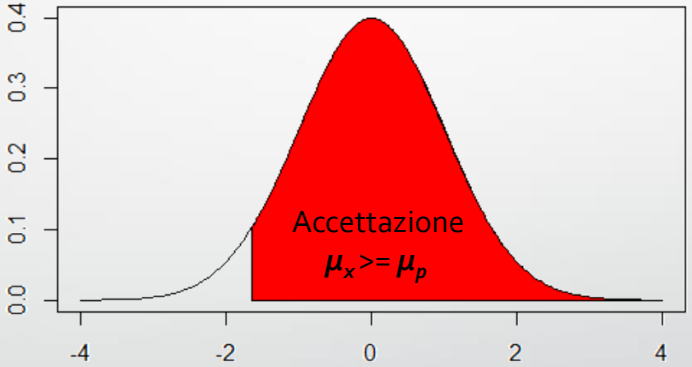


Verifica d'ipotesi sulla media 3/6

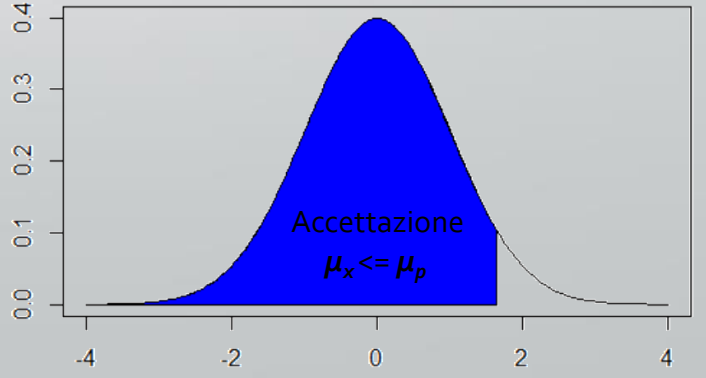
- *disponendo di un campione di numerosità limitata, si può affermare che la media μ_x della popolazione da cui esso è stato estratto è diversa dal valore prestabilito μ_p?*



- *è inferiore al valore prestabilito μ_p?*



- *è superiore al valore prestabilito μ_p?*



Verifica d'ipotesi sulla media 4/6

Es. *Supponiamo di aver rilevato il tempo medio di vita in h di un campione di 15 lampadine:*

x <- c(2928, 2997, 2689, 3081, 3011, 2996, 2962, 3007, 3000, 2953, 2792, 2947, 3094, 2913, 3017)

Per poter vendere queste lampadine occorre indicare sulla scatola il tempo medio di vita con un errore dell'1% se le vendo in Italia, e del 5% se le vendo all'estero.
La ditta di lampadine vende sia in Italia che all'estero e produce un solo tipo di scatola in cui è indicato 3010h come tempo medio di vita.
Verifichiamo se siamo confidenti nell'affermare che $\mu_x = \mu_p = 3010$*, ovvero che la ditta è in regola secondo le norme internazionali e locali.*

mean(x)
[1] 2959.133

Come si può notare la media campionaria è inferiore a quella dichiarata ($\mu_x < \mu_p$).
Vediamo se questa anomalia è dovuta al caso o se l'azienda non rispetta le norme.

Verifica d'ipotesi sulla media 5/6

Per fare questo dobbiamo effettuare un test t-di student ad una coda con la regione di rifiuto a sinistra ($\mu_x < \mu_p$):
Utilizziamo il comando: ***t.test****(x, mu=3010,* ***alternative="less")***

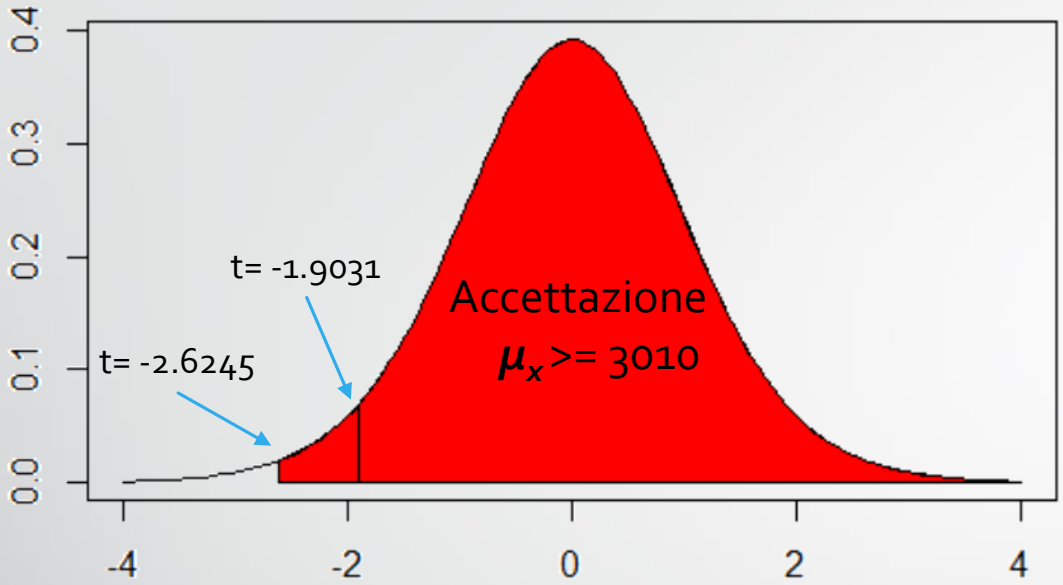
```
> t.test(x,mu=3010, alternative="less")

 One Sample t-test

data:  x
t = -1.9031, df = 14, p-value = 0.0389
alternative hypothesis: true mean is less than 3010
95 percent confidence interval:
     -Inf 3006.211
sample estimates:
mean of x
 2959.133
```

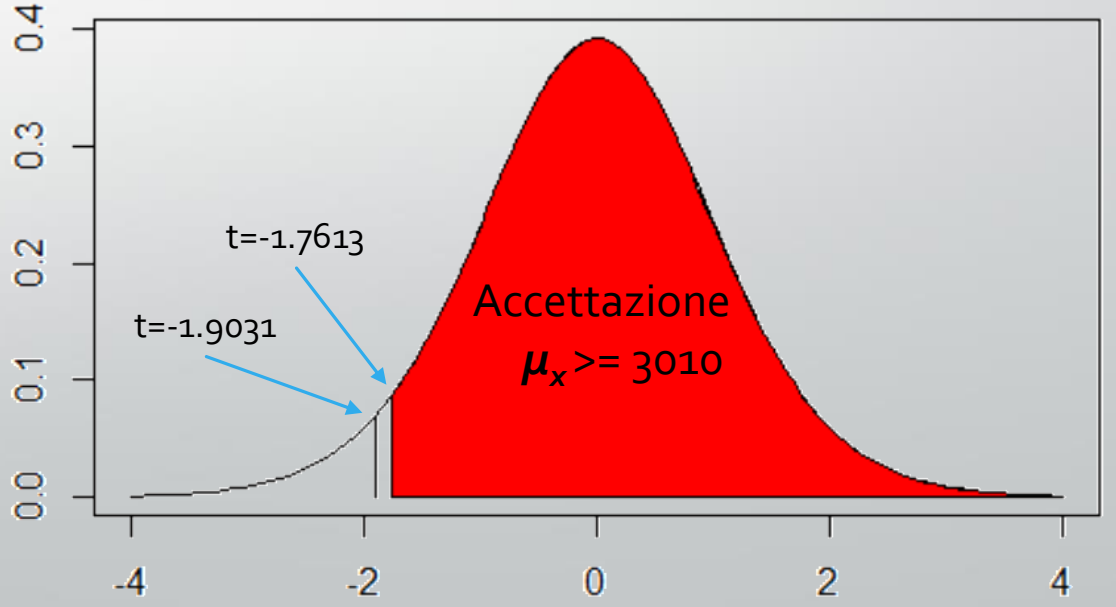
Fissato α=0.01 notiamo che α<p-value per cui l'azienda può vendere queste lampadine in Italia scrivendo 3010 come tempo medio di vita. Ma se fissiamo α=0.05 notiamo che α>p-value per cui l'Azienda non è conforme alle norme estere.

Verifica d'ipotesi sulla media 6/6



Fissato α=0.01 notiamo che α<p-value per cui l'azienda può vendere queste lampadine in Italia

Ma se fissiamo α=0.05 notiamo che α>p-value per cui l'Azienda non è conforme alle norme estere.



Codice Grafici

Confronto fra le medie di gruppi 1/2

Quando si vuole valutare la differenza tra le medie in due campioni, per vedere se fanno parte della stessa popolazione, si costruisce un test t come segue

$$t = \frac{\overline{x_1} - \overline{x_2}}{\bar{s}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad \text{dove} \quad \bar{s} = \sqrt{\frac{(n_1 - 1)\bar{s}_1^2 + (n_2 - 1)\bar{s}_2^2}{n_1 + n_2 - 2}}$$

Es. Supponiamo di avere due campioni di autovetture guidate nel primo gruppo da 5 uomini e nel secondo da 7 donne. Di questi due gruppi sono stati calcolati la spesa media per le riparazioni e il relativo scostamento medio campionario. Per semplicità generiamo casualmente da una normale i due gruppi impostando i valori calcolati

```
uomini <- rnorm(5,mean=540,sd=299)
donne <- rnorm(7,mean=300,sd=238)
```

Ci chiediamo se c'è differenza statisticamente significativa tra le medie dei due gruppi.

Confronto fra le medie di gruppi 2/2

t.test(uomini, donne, alternative="greater")

```
> t.test(uomini, donne, alternative="greater")

 Welch Two Sample t-test

data:  uomini and donne
t = 1.421, df = 9.735, p-value = 0.09328
alternative hypothesis: true difference in means is greater th
an 0
95 percent confidence interval:
 -79.9412      Inf
sample estimates:
mean of x mean of y
 571.8473  285.3752
```

Poiché il p-value è uguale al 9.3%, maggiore del 5%, si accetta l'ipotesi nulla di uguaglianza delle medie tra i due gruppi. Quindi uomini e donne sostengono la stessa spesa media per le riparazioni

Verifica di ipotesi di indipendenza

Per verificare se esiste indipendenza statistica tra due variabili X e Y si utilizza il test $\widetilde{X^2}$.

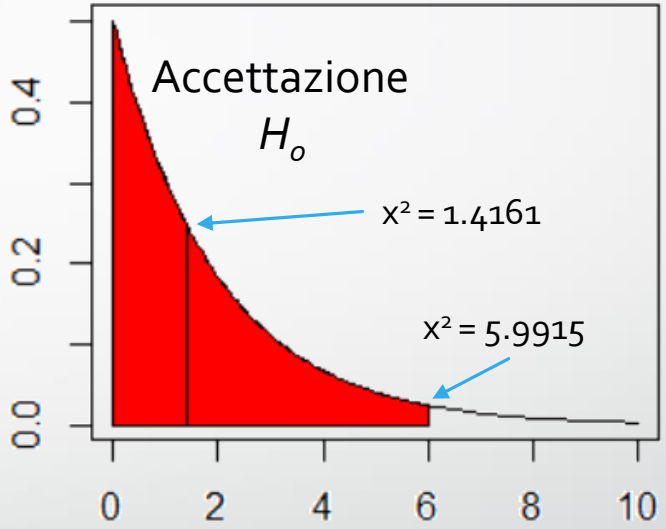
$\widetilde{X^2} = 0$ indica indipendenza statistica tra X e Y
$\widetilde{X^2} = 1$ indica massima connessione tra X e Y

Se $\widetilde{X^2}$ cade nell'intervallo (0,1) la connessione è tanto maggiore quanto l'indice di avvicina a 1.

Con il comando ***chisq.test(****tabella****)*** possiamo valutare la forza e la significatività statistica della connessione.

Verifica di ipotesi di indipendenza

```
> HRange<-cut(dataset$hinternet_lv, breaks=c(0, 5, 10, 15), include.lowest=TRUE);
> tabella<-table(dataset$domicilio,HRange)
> tabella
   HRange
     [0,5] (5,10] (10,15]
  0     45     23        1
  1     23     14        2
```



Codice grafico

```
> tabella<-table(dataset$domicilio,HRange)
> chisq.test(tabella)

 Pearson's Chi-squared test

data:  tabella
X-squared = 1.4161, df = 2, p-value = 0.4926
```

Dato che il P-Value è maggiore di 0.05, siamo portati a rifiutare una connessione tra le 2 variabili, supportando l'ipotesi H_0 di indipendenza.

Consegna

1. Una fabbrica di funi per arrampicata sportiva ha ottenuto i seguenti risultati espressi in Newton in 25 prove di rottura,
   per un nuovo tipo di funi:

   1975, 1869, 1879, 1790, 1860, 1895, 1810, 1831, 1759, 1585, 1553, 1774, 1640
   1761, 1946, 1915, 1894, 1971, 1876, 1716, 1652, 1591, 1700, 1842, 1781

   Sapendo che le funi tradizionali hanno una resistenza di rottura pari a 1730N, ci si chiede se il nuovo tipo abbia significativamente migliorato la qualità delle funi con una fiducia del 95%.

Consegna

2. Una famiglia consuma mediamente 100 litri di acqua al mese con una deviazione standard di 6 litri.
   La famiglia al piano di sopra ha un consumo medio di 120 litri al mese con una deviazione standard di 10 litri.
   Sapendo che il numero di componenti famigliari consumatori è identico, ci si chiede se le due famiglie hanno un consumo sproporzionato di acqua l'una rispetto all'altra con un livello di confidenza del 97%

3. Valutare se esiste indipendenza statistica tra:
   - *single* **vs** *hinternet_we*
   - *studio* **vs** *hlib_lv*

   Rappresentare tutti i test anche attraverso l'impiego dei grafici.