

Laboratorio di ST1 - Lezione 7

Antonietta di Salvatore

Dipartimento di Matematica
Università degli Studi Roma Tre

Outline

- ▶ Statistica descrittiva e grafici
- ▶ Verifica di ipotesi e intervalli di confidenza

II Dataset

Le variabili:

- ▶ Gender: Male or Female
- ▶ FSIQ: Full Scale IQ scores based on the four Wechsler (1981) subtests
- ▶ VIQ: Verbal IQ scores based on the four Wechsler (1981) subtests
- ▶ PIQ: Performance IQ scores based on the four Wechsler (1981) subtests
- ▶ Weight: body weight in pounds
- ▶ Height: height in inches
- ▶ MRI Count: total pixel Count from the 18 MRI scans

Carichiamo i dati

```
brain=read.table('brain.txt', header = TRUE, sep = ',', dec =  
'.'
```

Frequenze della variabile VIQ: assolute, relative, cumulate

```
> table(brain$VIQ)
 71  77  83  86  90  91  93  96 100 107
 1   1   2   2   5   1   1   4   1   1
112 114 120 123 126 129 132 136 145 150
 1   1   1   1   2   5   3   1   3   3
```

```
> VIQint=cut(brain$VIQ, breaks=c(70, 90, 110, 130, 150))
```

```
> table(VIQint)
```

```
VIQint
(70,90]  (90,110]  (110,130]  (130,150]
    11         8         11         10
```

```
> rel=table(VIQint)/length(brain$VIQ)
```

```
> rel
```

```
VIQint
(70,90]  (90,110]  (110,130]  (130,150]
 0.275   0.200   0.275   0.250
```

```
> cumsum(rel)
```

```
(70,90]  (90,110]  (110,130]  (130,150]
 0.275   0.475   0.750   1.000
```

Grafici 1: Istogramma

Costruiamo listogramma della variabile Weight (virtualmente continua) prima con le frequenze assolute e poi con quelle relative (basta aggiungere prob=T !):

```
hist(brain$Weight, col = 3, main='Istogramma di Weight')
```

```
hist(brain$Weight, col = 3, main='Istogramma di Weight', prob=TRUE)
```

Grafici 2: Box Plot

```
boxplot(brain$VIQ, col=3, main='Boxplot di VIQ')
```

Richiamo:

La linea centrale evidenzia la mediana

Sopra e sotto la mediana ci sono il primo e il terzo quartile

la lunghezza della scatola é detta intervallo o scarto interquartile (SIQ)

Le linee al termine dei *baffi* sono poste a $Q1 - 1.5SIQ$ (sotto) e $Q3 + 1.5SIQ$ (sopra)

Le osservazione esterne alle linee che terminano il baffo sono indicate da pallini (potenziali outliers)

Media

Osserviamo che nel dataset c' é qualche dato 'Non Disponibile' (Not Available: NA)
Se proviamo a calcolare la media della variabile 'Height'...

```
>mean(brain$Height)  
[1] NA
```

Possiamo escludere i dati NA attraverso il seguente comando: ridenominiamo Height senza NA come H2, e poi ne potremo calcolare la media

```
>H2=na.exclude(brain$Height)  
>mean(H2)
```

Mediana

La Mediana si trova semplicemente con il comando 'median':

```
> median(brain$PIQ)  
[1] 115
```

la mediana é la modalit  che divide la popolazione in due parti di uguale numerosit 

I Quartili

I Quartili dividono la popolazione in quattro parti di uguale numerosità
il primo ha il 25% della popolazione a sinistra ed il 75% a destra
il secondo coincide con la mediana e ha il 50% della popolazione sia a sinistra che a destra
il terzo ha il 75% della popolazione a sinistra ed il 25% a destra
il quarto ha il 100% della popolazione a sinistra.
Possiamo averli tutti con un solo comando:

```
>quantile(brain$MRICount)
```

0%	25%	50%	75%	100%
790619.0	855918.5	905399.0	950078.0	1079549.0

Possiamo calcolare gli scarti interquartili, cioè le differenze tra un quartile e l'altro

```
>quantile(brain$MRICount, 1)-quantile(brain$MRICount, 0.75)  
100%  
129471
```

```
>quantile(brain$MRICount, 0.75)-quantile(brain$MRICount, 0.5)  
75%  
44679
```

```
>quantile(brain$MRICount, 0.5)-quantile(brain$MRICount, 0.25)  
50%  
49480.5
```

Summary

Il comando Summary ci fornisce molte misure, tutte insieme...

```
> summary(brain$VIQ)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
71.0	90.0	113.0	112.3	129.8	150.0

```
> summary(brain)
```

Scatterplots

```
plot(brain)
```

Esercizio 1.

Per la generica voce di un inventario un'impresa mercantile sia X la variabile casuale *valore inventariato-valore certificato*. Un certificatore contabile estrae a sorte un campione di 13 voci:

17.60, -150.84, 203.84, 76.06, 47.86, -162.26, 310.77, 224.92, -11.05, -111.16 , 28.66 , 103.58, -146.34.

Si sottoponga a verifica al 95% l'ipotesi che $\text{media}=0$ contro l'alternativa $\text{media}>0$, ossia che l'inventario é gonfiato.

```
x=c(17.60, -150.84, 203.84, 76.06, 47.86, -162.26, 310.77,  
224.92, -11.05, -111.16 , 28.66 , 103.58, -146.34)  
t.test(x, mu = 0, alternative = 'greater')
```

Esercizio 2.

Il direttore dell'Ospedale ABC, situato in un quartiere molto povero nei sobborghi di New York, sospetta che i neonati che nascono lí abbiano un peso inferiore rispetto alla media nazionale (pari a 3.2 Kg), tale da dover richiedere un intervento di prevenzione sulla malnutrizione delle donne del quartiere. Il direttore decide allora di verificare l'ipotesi $H_0 : \mu = 3.2$ in alternativa all'ipotesi $H_1 : \mu < 3.2$. Il direttore misura, quindi, il peso $(x_1, x_2, \dots, x_{20})$ di 20 bambini scelti casualmente tra i neonati nell'ultimo anno e riscontra che i loro pesi sono pari a 2.2, 3.5, 3.0, 1.1, 2.4, 4.2, 3.4, 2.1, 1.7, 0.8, 3.1, 2.0, 2.5, 3.1, 1.9, 3.4, 2.8, 3.5, 3.4, 1.0. Si verifichi l'ipotesi che la media del quartiere é uguale a quella di New York contro l'ipotesi di malnutrizione al livello del 5%.

Esercizio 3.

In un esame di psicologia vengono misurati i tempi di reazione su un certo numero di individui, riscontrando un tempo medio di 1 secondo. Da studi pregressi, lo scarto quadratico σ é noto essere pari a 0.05 secondi. Quale deve essere il numero minimo di osservazioni campionarie n per avere un'ampiezza dell'intervallo pari al piú a 0.02 secondi ed un intervallo di confidenza pari al 99%?

Soluzione In generale, all'aumentare di n l'intervallo di confidenza si restringe. Si ha:

$$n \geq (2z_{1-\alpha/2}\sigma/L)^2$$

con L ampiezza dell'intervallo

$$a = 1 - 0.01/2$$

$$q = \text{qnorm}(a)$$

$$n1 = (2 * q * (0.05/0.02))^2$$

Esercizio 4.

Negli Stati Uniti la clausola *Union Shop* in un contratto di lavoro obbliga il lavoratore a iscriversi a un sindacato non appena il contratto é firmato. Nel 1973 31 stati avevano accettato la *Union Shop*, mentre 19 stati avevano accettato le leggi *Right-to-Work*, che aboliscono di fatto la *Union Shop*. Un campione di 5 stati da ogni gruppo ha mostrato i seguenti salari medi orari in ogni stato (in dollari):

Stati con *Union Shop* :4.00, 3.10 , 3.60, 4.20, 4.60

Stati con la *Right-to-Work*. 3.50, 3.60, 3.20, 3.90, 2.80

Sulla base di queste cifre, si 'e detto che le leggi *Right-to-Work* costano ai lavoratori 50 centesimi lora. Questa affermazione é corretta? Costruire l'intervallo di confidenza al 95%. (-0,25; 1,25). Considerare i casi in cui la varianza tra le due popolazioni sia uguale o diversa.