Statistica Inferenziale

a) L'Intervallo di Confidenza
b) La distribuzione t di Student
c) La differenza delle medie
d) L'intervallo di confidenza della differenza

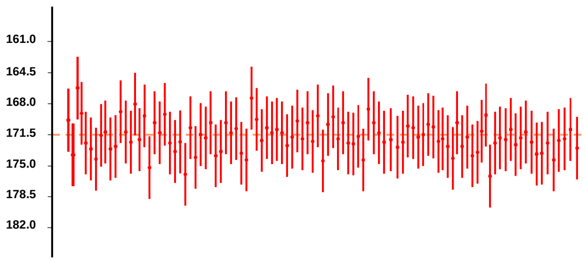
Stimare i Parametri della Popolazione

- La media del gruppo (campione) è una *stima puntuale* del parametro della popolazione
- Ogni media di gruppo fornisce una diversa stima connessa alle *fluttuazioni casuali* dovute al campionamento
- La stima puntuale non da indicazioni sulla variabilità della stima
- Costruisco un intervallo centrato intorno alla media di gruppo sul quale ho una certa *confidenza* che il parametro della popolazione cada nell'intervallo
- L'intervallo di confidenza è la *stima intervallare* del parametro della popolazione

Intervallo di Confidenza e Parametro

Intervallo di Confidenza e Parametro

Intervallo di Confidenza e Parametro

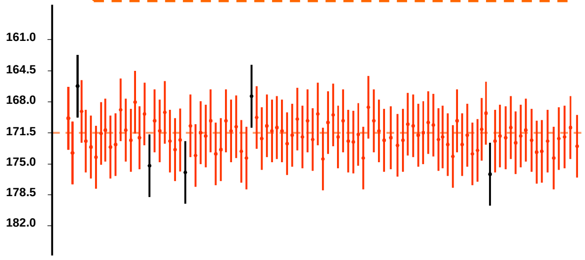


Intervallo di Confidenza

- Gli intervalli di confidenza sono definiti come un intervallo di valori costruito a partire dai dati
- All'interno dell'intervallo ho una certa probabilità (tipicamente 95%) che sia compreso il parametro della popolazione

Intervallo di Confidenza e Parametro

Nel 95% circa dei campioni possibili l'intervallo di confidenza al 95% comprende il parametro della popolazione (171.5 cm)



Intervallo di Confidenza

- Gli intervalli di confidenza sono definiti come un intervallo di valori costruito a partire dai dati
- All'interno dell'intervallo ho una certa probabilità (tipicamente 95%) che sia compreso il parametro della popolazione
- Per calcolare l'intervallo utilizzo le proprietà della distribuzione di campionamento delle medie

Esempio di Calcolo dell'Intervallo di Confidenza al 95%

$$\left(\overline{X} - z \cdot \sigma / \sqrt{n}, \overline{X} + z \cdot \sigma / \sqrt{n}\right)$$

Informazioni

- $n = 20$
- $\overline{x} = 170$
- $\sigma = 8.5$
- $z = 1.96$

Limite Inferiore

170 – 1.96*8.5/√ 20 = 170 – 3.72 = 166.28

Limite Superiore

170 + 1.96*8.5/√ 20 = 170 + 3.72 = 173.72

Intervallo di Confidenza

Proprietà

- Maggiore è l'ampiezza dell' Intervallo di Confidenza minore è la precisione della stima
- La sua ampiezza, e quindi la precisione della stima, varia con la numerosità dello studio e il grado di confidenza desiderato
  - All'aumentare della numerosità l'ampiezza diminuisce e la precisione aumenta
  - All'aumentare del grado di confidenza (es. 99% invece di 95%) l'ampiezza aumenta e la precisione diminuisce

Se σ è sconosciuta ?

Problema

Se la varianza della popolazione σ^2 non è nota ?

(NB se μ non è nota, è probabile che anche σ^2 non sia nota)

Soluzione

Utilizzo la varianza campionaria s^2 **come stima di** σ^2

(NB nella formula della varianza divido per *(n-1)***: i gradi di libertà)**

La distribuzione t di student

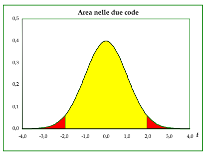
$$\boxed{Z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}}} \longrightarrow \boxed{t = \frac{\overline{X} - \mu}{s / \sqrt{n}}}$$

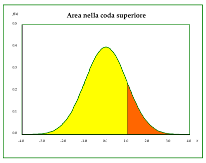
Il nuovo rapporto standardizzato non ha una distribuzione normale standardizzata perché devo tener conto anche della variabilità di *s* **che sarà maggiore quando** *n* **è piccolo.**

Questo rapporto è distribuito come una *t* **di student con** *n-1* **gradi di libertà**

Percentili della distribuzione t di Student

	PROBABILITA' (2 code)				PROBABILITA' (1 coda)			
GL	0,1	0,05	0,02	0,01	0,05	0,025	0,01	0,00 5
1	6,31	12,71	31,82	63,66	6,31	12,71	31,82	63,66
2	2,92	4,30	6,96	9,92	2,92	4,30	6,96	9,92
3	2,35	3,18	4,54	5,84	2,35	3,18	4,54	5,84
4	2,13	2,78	3,75	4,60	2,13	2,78	3,75	4,60
5	2,02	2,57	3,36	4,03	2,02	2,57	3,36	4,03
6	1,94	2,45	3,14	3,71	1,94	2,45	3,14	3,71
7	1,89	2,36	3,00	3,50	1,89	2,36	3,00	3,50
8	1,86	2,31	2,90	3,36	1,86	2,31	2,90	3,36
9	1,83	2,26	2,82	3,25	1,83	2,26	2,82	3,25
10	1,81	2,23	2,76	3,17	1,81	2,23	2,76	3,17
11	1,80	2,20	2,72	3,11	1,80	2,20	2,72	3,11
12	1,78	2,18	2,68	3,05	1,78	2,18	2,68	3,05
13	1,77	2,16	2,65	3,01	1,77	2,16	2,65	3,01
14	1,76	2,14	2,62	2,98	1,76	2,14	2,62	2,98
15	1,75	2,13	2,60	2,95	1,75	2,13	2,60	2,95
16	1,75	2,12	2,58	2,92	1,75	2,12	2,58	2,92
17	1,74	2,11	2,57	2,90	1,74	2,11	2,57	2,90
18	1,73	2,10	2,55	2,88	1,73	2,10	2,55	2,88
19	1,73	2,09	2,54	2,86	1,73	2,09	2,54	2,86
20	1,72	2,09	2,53	2,85	1,72	2,09	2,53	2,85
21	1,72	2,08	2,52	2,83	1,72	2,08	2,52	2,83
22	1,72	2,07	2,51	2,82	1,72	2,07	2,51	2,82
23	1,71	2,07	2,50	2,81	1,71	2,07	2,50	2,81
24	1,71	2,06	2,49	2,80	1,71	2,06	2,49	2,80
25	1,71	2,06	2,49	2,79	1,71	2,06	2,49	2,79
26	1,71	2,06	2,48	2,78	1,71	2,06	2,48	2,78
27	1,70	2,05	2,47	2,77	1,70	2,05	2,47	2,77
28	1,70	2,05	2,47	2,76	1,70	2,05	2,47	2,76
29	1,70	2,05	2,46	2,76	1,70	2,05	2,46	2,76
30	1,70	2,04	2,46	2,75	1,70	2,04	2,46	2,75
∞	1,64	1,96	2,05	2,33	1,64	1,96	2,05	2,33





La distribuzione *t* di student

Caratteristiche

1. **È una distribuzione continua**
2. **È simmetrica rispetto alla media: μ**
3. **Media , mediana e moda coincidono**
4. **È una distribuzione di *probabilità***
5. **Se *n* è basso i valori nelle code sono più probabili**
6. **Al crescere di *n* la distribuzione approssima la *gaussiana standardizzata***

Distribuzione *t* di Student e Intervallo di Confidenza

Consideriamo i dati sull'altezza raccolti da un gruppo di studenti

n = 20

$\bar{x}$ = 172.0

***s* = 10.0**

Qual è l'intervallo di confidenza al 95% della media ?

Distribuzione *t* di Student e Intervallo di Confidenza

Occorre modificare la formula precedente

$$\Pr\left\{ \overline{X} - 1.96 * \sigma/\sqrt{n} \le \mu \le \overline{X} + 1.96 * \sigma/\sqrt{n} \right\} = 0.95$$

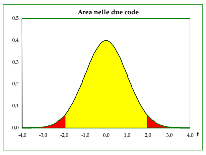
tenendo conto delle nuove informazioni

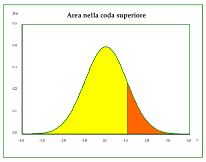
$$\Pr\left\{ \overline{X} - t_{n-1} * S/\sqrt{n} \le \mu \le \overline{X} + t_{n-1} * S/\sqrt{n} \right\} = 0.95$$

Quali valori della distribuzione *t* di Student con 19 gradi di libertà lasciano un'area nelle due code pari a 0.05 ?

Percentili della distribuzione t di Student

	PROBABILITA' (2 code)				PROBABILITA' (1 coda)			
GL	0,1	0,05	0,02	0,01	0,05	0,025	0,01	0,00 5
1	6,31	12,71	31,82	63,66	6,31	12,71	31,82	63,66
2	2,92	4,30	6,96	9,92	2,92	4,30	6,96	9,92
3	2,35	3,18	4,54	5,84	2,35	3,18	4,54	5,84
4	2,13	2,78	3,75	4,60	2,13	2,78	3,75	4,60
5	2,02	2,57	3,36	4,03	2,02	2,57	3,36	4,03
6	1,94	2,45	3,14	3,71	1,94	2,45	3,14	3,71
7	1,89	2,36	3,00	3,50	1,89	2,36	3,00	3,50
8	1,86	2,31	2,90	3,36	1,86	2,31	2,90	3,36
9	1,83	2,26	2,82	3,25	1,83	2,26	2,82	3,25
10	1,81	2,23	2,76	3,17	1,81	2,23	2,76	3,17
11	1,80	2,20	2,72	3,11	1,80	2,20	2,72	3,11
12	1,78	2,18	2,68	3,05	1,78	2,18	2,68	3,05
13	1,77	2,16	2,65	3,01	1,77	2,16	2,65	3,01
14	1,76	2,14	2,62	2,98	1,76	2,14	2,62	2,98
15	1,75	2,13	2,60	2,95	1,75	2,13	2,60	2,95
16	1,75	2,12	2,58	2,92	1,75	2,12	2,58	2,92
17	1,74	2,11	2,57	2,90	1,74	2,11	2,57	2,90
18	1,73	2,10	2,55	2,88	1,73	2,10	2,55	2,88
19	1,73	2,09	2,54	2,86	1,73	2,09	2,54	2,86
20	1,72	2,09	2,53	2,85	1,72	2,09	2,53	2,85
21	1,72	2,08	2,52	2,83	1,72	2,08	2,52	2,83
22	1,72	2,07	2,51	2,82	1,72	2,07	2,51	2,82
23	1,71	2,07	2,50	2,81	1,71	2,07	2,50	2,81
24	1,71	2,06	2,49	2,80	1,71	2,06	2,49	2,80
25	1,71	2,06	2,49	2,79	1,71	2,06	2,49	2,79
26	1,71	2,06	2,48	2,78	1,71	2,06	2,48	2,78
27	1,70	2,05	2,47	2,77	1,70	2,05	2,47	2,77
28	1,70	2,05	2,47	2,76	1,70	2,05	2,47	2,76
29	1,70	2,05	2,46	2,76	1,70	2,05	2,46	2,76
30	1,70	2,04	2,46	2,75	1,70	2,04	2,46	2,75
∞	1,64	1,96	2,05	2,33	1,64	1,96	2,05	2,33





Calcolo dell'Intervallo di Confidenza

Inseriamo le informazioni raccolte nella formula

$$\left(\overline{X} - t_{n-1} \cdot s/\sqrt{n}, \overline{X} + t_{n-1} \cdot s/\sqrt{n} \right)$$

$$\left(172 - 2.09 \cdot 10/\sqrt{20}, 172 + 2.09 \cdot 10/\sqrt{20} \right)$$

n = 20

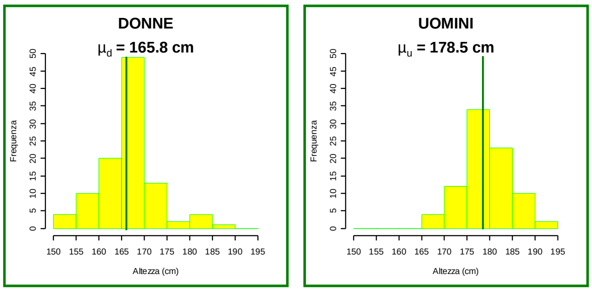
$\overline{x}$ = 172.0

***s* = 10.0**

t = 2.09

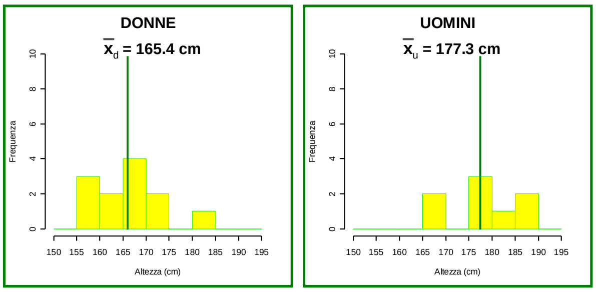
I limiti dell'intervallo di confidenza sono 167.33 e 176.33

Altezza della Popolazione di Studenti per Genere



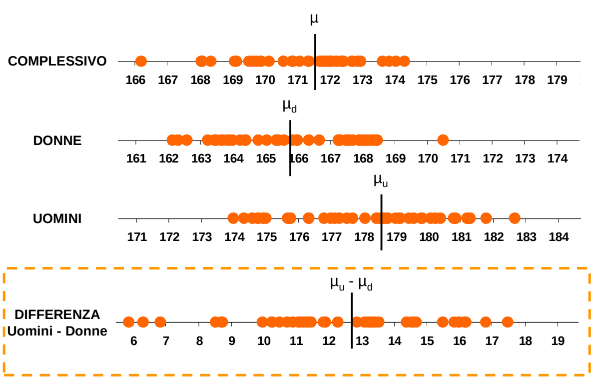
La differenza dell'altezza tra Uomini e Donne: $\mu_u - \mu_d = 12.7$ cm

Altezza di un Campione di Studenti per Genere



La differenza dell'altezza tra Uomini e Donne: $\bar{x}_u - \bar{x}_d = 11.8$ cm

Distribuzione delle Medie Campionarie dell'Altezza



Distribuzione delle Medie Campionarie

Caratteristiche della distribuzione delle medie campionarie

1. **È approssimativamente Gaussiana**
2. **La media della distribuzione è μ**
3. **La deviazione standard della distribuzione è uguale a $\sigma/\sqrt{n}$**

E la distribuzione della differenza delle medie campionarie?

Occorre distinguere in due casi differenti:

- ✓ **Campioni Indipendenti (es. Uomini-Donne)**
- ✓ **Campioni Appaiati (es. 2 misure ripetute)**

Campioni Indipendenti

Distribuzione della differenza delle medie campionarie

1. **È approssimativamente Gaussiana**
2. **La media della distribuzione è $\mu_1 - \mu_2$**
3. **L'errore standard della distribuzione è uguale a:**

$$\sigma\sqrt{\left(\frac{1}{n_1}\right)+\left(\frac{1}{n_2}\right)}$$ ⟵ **La varianza delle due popolazioni è uguale**

Di questa situazione non ci occupiamo

$$\sqrt{\left(\frac{\sigma_1^2}{n_1}\right)+\left(\frac{\sigma_2^2}{n_2}\right)}$$ ⟵ **La varianza delle due popolazioni non è uguale**

Distribuzione della differenza delle medie

Problema

Qual è la probabilità che la differenza media di altezza tra uomini e donne in un gruppo formato da 10 uomini e 10 donne sia inferiore a 6 cm?

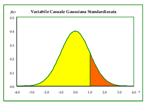
Soluzione

La distribuzione della differenza delle medie campionarie è gaussiana allora utilizzo il rapporto standardizzato:

$$Z=\frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}=\frac{6-12.7}{8.5\sqrt{\frac{1}{10}+\frac{1}{10}}}=\frac{-6.7}{3.8}=-1.76$$

La tavola della distribuzione Gaussiana Standardizzata

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.500	0.496	0.492	0.488	0.484	0.480	0.476	0.472	0.468	0.464
0.1	0.460	0.456	0.452	0.448	0.444	0.440	0.436	0.433	0.429	0.425
0.2	0.421	0.417	0.413	0.409	0.405	0.401	0.397	0.394	0.390	0.386
0.3	0.382	0.378	0.374	0.371	0.367	0.363	0.359	0.356	0.352	0.348
0.4	0.345	0.341	0.337	0.334	0.330	0.326	0.323	0.319	0.316	0.312
0.5	0.309	0.305	0.302	0.298	0.295	0.291	0.288	0.284	0.281	0.278
0.6	0.274	0.271	0.268	0.264	0.261	0.258	0.255	0.251	0.248	0.245
0.7	0.242	0.239	0.236	0.233	0.230	0.227	0.224	0.221	0.218	0.215
0.8	0.212	0.209	0.206	0.203	0.200	0.198	0.195	0.192	0.189	0.187
0.9	0.184	0.181	0.179	0.176	0.174	0.171	0.169	0.166	0.164	0.161
1.0	0.159	0.156	0.154	0.152	0.149	0.147	0.145	0.142	0.140	0.138
1.1	0.136	0.133	0.131	0.129	0.127	0.125	0.123	0.121	0.119	0.117
1.2	0.115	0.113	0.111	0.109	0.107	0.106	0.104	0.102	0.100	0.099
1.3	0.097	0.095	0.093	0.092	0.090	0.089	0.087	0.085	0.084	0.082
1.4	0.081	0.079	0.078	0.076	0.075	0.074	0.072	0.071	0.069	0.068
1.5	0.067	0.066	0.064	0.063	0.062	0.061	0.059	0.058	0.057	0.056
1.6	0.055	0.054	0.053	0.052	0.051	0.049	0.048	0.048	0.046	0.046
1.7	0.045	0.044	0.043	0.042	0.041	0.040	**0.039**	0.038	0.037	0.037
1.8	0.036	0.035	0.034	0.034	0.033	0.032	0.031	0.030	0.029	0.029
1.9	0.029	0.028	0.027	0.027	0.026	0.026	0.025	0.024	0.024	0.023
2.0	0.023	0.022	0.022	0.021	0.021	0.020	0.020	0.019	0.019	0.018
2.1	0.018	0.017	0.017	0.017	0.016	0.016	0.015	0.015	0.015	0.014
2.2	0.014	0.014	0.013	0.013	0.013	0.012	0.012	0.012	0.011	0.011
2.3	0.011	0.010	0.010	0.010	0.010	0.009	0.009	0.009	0.009	0.008
2.4	0.008	0.008	0.008	0.008	0.007	0.007	0.007	0.007	0.007	0.006
2.5	0.006	0.006	0.006	0.006	0.006	0.005	0.005	0.005	0.005	0.005
2.6	0.005	0.005	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004
2.7	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.003
2.8	0.003	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002
2.9	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.001	0.001	0.001



$P(\bar{x}_u - \bar{x}_d < 6) =$

0.039

Calcolo dell'Intervallo di Confidenza al 95%

$$\left(\left(\bar{X}_u - \bar{X}_d\right) - z \cdot \sigma \cdot \sqrt{\left(\frac{1}{n_u}\right) + \left(\frac{1}{n_d}\right)}, \left(\bar{X}_u - \bar{X}_d\right) + z \cdot \sigma \cdot \sqrt{\left(\frac{1}{n_u}\right) + \left(\frac{1}{n_d}\right)}\right)$$

Informazioni

$n_u = 8$

$n_d = 12$

$\bar{x}_u = 177.3$

$\bar{x}_d = 165.4$

$\sigma = 8.5$

$z = 1.96$

Limite Inferiore

$$(177.3 - 165.4) - 1.96 \cdot 8.5 \cdot \sqrt{\left(\frac{1}{8}\right) + \left(\frac{1}{12}\right)} = 11.9 - 7.6 = 4.3$$

Limite Superiore

$$(177.3 - 165.4) + 1.96 \cdot 8.5 \cdot \sqrt{\left(\frac{1}{8}\right) + \left(\frac{1}{12}\right)} = 11.9 + 7.6 = 19.5$$

Se σ è sconosciuta ?

Problema

Se la varianza della popolazione σ^2 non è nota ?

Soluzione

Utilizzo le due varianze campionarie s_1^2 e s_2^2 per stimare la varianza della popolazione:

$$S^2_{pooled} = \frac{(n_1 - 1) \cdot S_1^2 + (n_2 - 1) \cdot S_2^2}{n_1 + n_2 - 2}$$

In questo caso la distribuzione delle differenza delle medie approssima alla distribuzione *t* con $n_1 + n_2 - 2$ gradi di libertà

Calcolo dell'Intervallo di Confidenza al 95%

$$\left(\left(\overline{X}_u - \overline{X}_d\right) - t_{gl} \cdot s_{pooled} \cdot \sqrt{\left(\frac{1}{n_u}\right) + \left(\frac{1}{n_d}\right)}, \left(\overline{X}_u - \overline{X}_d\right) + t_{gl} \cdot s_{pooled} \cdot \sqrt{\left(\frac{1}{n_u}\right) + \left(\frac{1}{n_d}\right)} \right)$$

Informazioni

$n_u = 8$

$n_d = 12$

$\overline{x}_u = 177.3$

$\overline{x}_d = 165.4$

$s_u^2 = 58.8$

$s_d^2 = 51.5$

$gl = 18$

$t_{18} = 2.10$

Limite Inferiore

$$(177.3 - 165.4) - 2.10 \cdot 7.4 \cdot \sqrt{\left(\frac{1}{8}\right) + \left(\frac{1}{12}\right)} = 11.9 - 7.1 = 4.8$$

Limite Superiore

$$(177.3 - 165.4) + 2.10 \cdot 7.4 \cdot \sqrt{\left(\frac{1}{8}\right) + \left(\frac{1}{12}\right)} = 11.9 + 7.1 = 19.0$$