Esercitazioni di statistica

Distribuzioni campionarie

Stefania Spina

Universitá di Napoli Federico II

stefania.spina@unina.it

3 Dicembre 2014

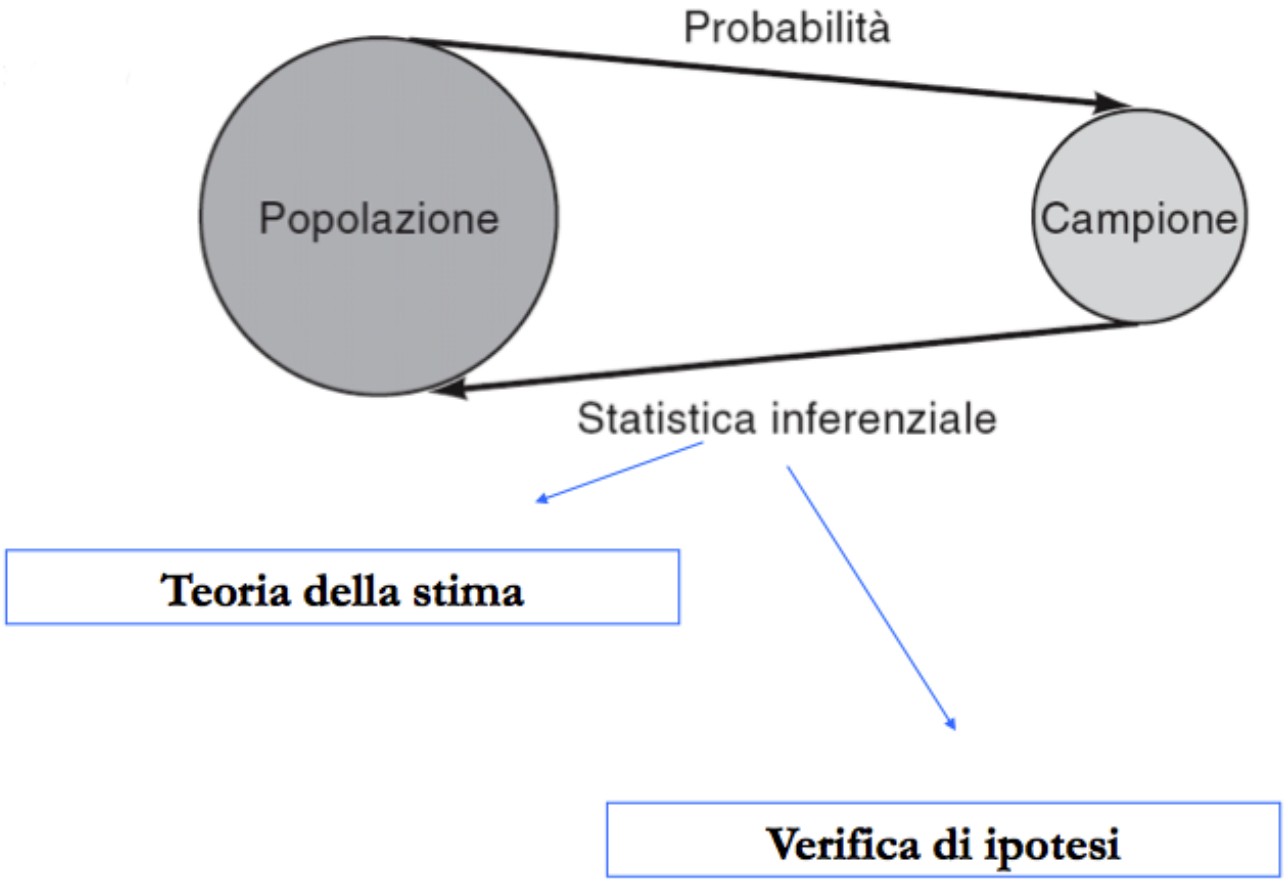
Lo studio della Statistica: come si articola

- Gli strumenti della statistica descrittiva permettono di sintetizzare e descrivere un insieme di dati
- La teoria della probabilità è usata per modellare l'incertezza e la variabilità della realtà, per poi poterla misurare e controllare (non eliminare).

La statistica inferenziale

Le tecniche della statistica inferenziale, sulla base del calcolo delle probabilità che permettono di calcolare la verosimiglianza di osservare o selezionare un particolare campione da una certa popolazione, consentono di trarre inferenze o conclusioni sulla popolazione a partire dal campione

La statistica inferenziale



Introduzione

Nell'inferenza statistica si possono, schematicamente, individuare due momenti distinti:

- Il momento della stima di una quantità statistica;
- Il momento della valutazione delle proprietà della quantità statistica stimata.

Di seguito l'attenzione è rivolta ad alcuni semplici metodi analitici utilizzabili per derivare la distribuzione di alcuni stimatori come quelli della media e della varianza.

Distribuzione di probabilità della media campionaria

Data la v.c. X con media e varianza finite, supponiamo che σ^2 sia nota. Uno stimatore accettabile di μ è la media campionaria

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad per\ \mu$$

ove le X_i sono, per costruzione, indipendenti oltre al fatto che sappiamo

$$E(\overline{X}) = \mu \qquad Var(\overline{X}) = \frac{\sigma^2}{n}$$

che ci permettono di costruire la v.c. standardizzata

$$Z = \frac{\overline{X} - \mu}{\sqrt{var(\overline{X})}} = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

Distribuzione di probabilità della media campionaria

Applicando il teorema del limite centrale si ha:

$$Z \xrightarrow{L} N(0,1)$$

Questo vuole dire che se n è sufficientemente grande, in pratica basta che sia $n \geq 30$, qualsiasi sia la distribuzione di X risulta

$$\overline{X} \approx N\left(\mu; \frac{\sigma^2}{n}\right)$$

Nel caso particolare in cui è $X \sim N(\mu, \sigma^2)$ segue immediatamente che, per una delle proprietà delle v.c. normali (una trasformazione lineare di normali indipendenti è ancora una normale) risulta

$$\overline{X} \sim N\left(\mu; \frac{\sigma^2}{n}\right)$$

qualsiasi sia n.

Esercizio 1

La media e la deviazione standard campionarie dei punteggi ottenuti dagli studenti all'ultimo anno Scholastic Aptitude Test (SAT) sono state rispettivamente 517 e 120.
Determina la probabilità approssimata che un campione casuale di 144 studenti ottenga un punteggio medio superiore a:

a. 507
b. 517
c. 537
d. 550

Soluzione esercizio 1

a.

$\overline{X} = 507$

$\mu = 517$

$\sigma = 120$

$n = 144$

$$P(\overline{X} > 507) = P\left(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} > \frac{507-517}{120/\sqrt{144}}\right) = P(z > -1) = 0.8413$$

Soluzione esercizio 1

b.

$\overline{X} = 517$

$\mu = 517$

$\sigma = 120$

$n = 144$

$$P(\overline{X} > 517) = P\left(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} > \frac{517-517}{120/\sqrt{144}}\right) = P(z > 0) = 0.50$$

Soluzione esercizio 1

c.

$\overline{X} = 537$

$\mu = 517$

$\sigma = 120$

$n = 144$

$$P(\overline{X} > 537) = P\left(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} > \frac{537-517}{120/\sqrt{144}}\right) = P(z > 2) = 0.0228$$

Soluzione esercizio 1

d.
$\overline{X} = 550$
$\mu = 517$
$\sigma = 120$
$n = 144$
$P(\overline{X} > 550) = P\left(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} > \frac{550-517}{120/\sqrt{144}}\right) = P(z > 3.3) = 0.50 - 0.4996 =$
0.0004

Esercizio 2

Una popolazione formata da operai maschi, presenta dei pesi corporei (in libbre) di media 167 e scarto quadratico medio 27.
Se si seleziona un campione di 36 elementi:

a) Quanto vale circa la probabilità che la media campionaria dei loro pesi stia tra 163 e 171?

b) E se si selezionano 144 operai?

Soluzione esercizio 2

a) Quanto vale circa la probabilità che la media campionaria dei loro pesi stia tra 163 e 171?

$$X \sim ?(167; 27^2)$$

$$n = 36$$

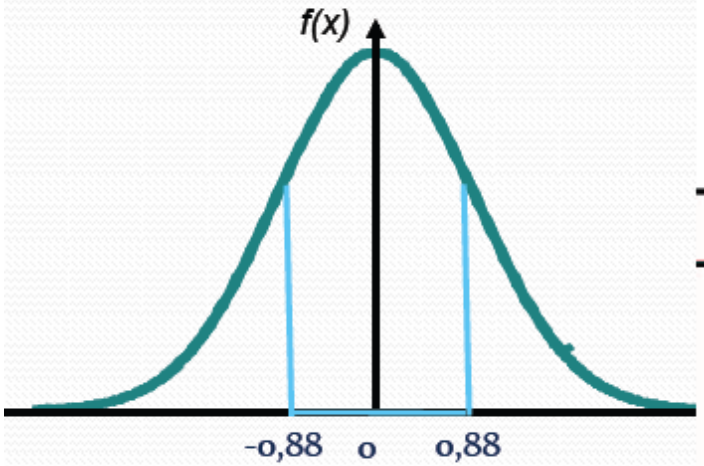
$$P(163 \leq \overline{X} \leq 171) = ?$$

$$\overline{X} \sim N\left(\mu; \frac{\sigma^2}{n}\right) \rightarrow \textit{per il teorema del limite centrale}$$

$$P(163 \leq \overline{X} \leq 171) = P\left(\frac{163 - 167}{\frac{27}{\sqrt{36}}} \leq Z \leq \frac{171 - 167}{\frac{27}{\sqrt{36}}}\right) =$$

$$= P(-0.88 \leq z \leq 0.88) \cong 0.63$$

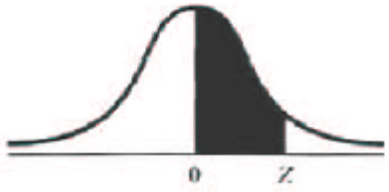
Soluzione esercizio 2



$$P(-0{,}88 \le z \le 0{,}88) = 2 * P(0 \le z \le 0{,}88)$$

$$\approx 2 * 0{,}3106 \approx 0.63$$

Tavola della distribuzione Normale Standardizzata



Area sottesa alla curva di densità normale standardizzata calcolata tra 0 e Z

Z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,0000	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319	0,0359
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,0753
0,2	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026	0,1064	0,1103	0,1141
0,3	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	0,1406	0,1443	0,1480	0,1517
0,4	0,1554	0,1591	0,1628	0,1664	0,1700	0,1736	0,1772	0,1808	0,1844	0,1879
0,5	0,1915	0,1950	0,1985	0,2019	0,2054	0,2088	0,2123	0,2157	0,2190	0,2224
0,6	0,2257	0,2291	0,2324	0,2357	0,2389	0,2422	0,2454	0,2486	0,2517	0,2549
0,7	0,2580	0,2611	0,2642	0,2673	0,2704	0,2734	0,2764	0,2794	0,2823	0,2852
0,8	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0,3106	0,3133
0,9	0,3159	0,3186	0,3212	0,3238	0,3264	0,3289	0,3315	0,3340	0,3365	0,3389
1,0	0,3413	0,3438	0,3461	0,3485	0,3508	0,3531	0,3554	0,3577	0,3599	0,3621
1,1	0,3643	0,3665	0,3686	0,3708	0,3729	0,3749	0,3770	0,3790	0,3810	0,3830
1,2	0,3849	0,3869	0,3888	0,3907	0,3925	0,3944	0,3962	0,3980	0,3997	0,4015
1,3	0,4032	0,4049	0,4066	0,4082	0,4099	0,4115	0,4131	0,4147	0,4162	0,4177
1,4	0,4192	0,4207	0,4222	0,4236	0,4251	0,4265	0,4279	0,4292	0,4306	0,4319
1,5	0,4332	0,4345	0,4357	0,4370	0,4382	0,4394	0,4406	0,4418	0,4429	0,4441
1,6	0,4452	0,4463	0,4474	0,4484	0,4495	0,4505	0,4515	0,4525	0,4535	0,4545
1,7	0,4554	0,4564	0,4573	0,4582	0,4591	0,4599	0,4608	0,4616	0,4625	0,4633
1,8	0,4641	0,4649	0,4656	0,4664	0,4671	0,4678	0,4686	0,4693	0,4699	0,4706
1,9	0,4713	0,4719	0,4726	0,4732	0,4738	0,4744	0,4750	0,4756	0,4761	0,4767
2,0	0,4772	0,4778	0,4783	0,4788	0,4793	0,4798	0,4803	0,4808	0,4812	0,4817
2,1	0,4821	0,4826	0,4830	0,4834	0,4838	0,4842	0,4846	0,4850	0,4854	0,4857
2,2	0,4861	0,4864	0,4868	0,4871	0,4875	0,4878	0,4881	0,4884	0,4887	0,4890
2,3	0,4893	0,4896	0,4898	0,4901	0,4904	0,4906	0,4909	0,4911	0,4913	0,4916
2,4	0,4918	0,4920	0,4922	0,4925	0,4927	0,4929	0,4931	0,4932	0,4934	0,4936

Soluzione esercizio 2

b) E se si selezionano 144 operai?

$$X \sim N\left(\mu; \frac{\sigma^2}{n}\right) = \left(167; \frac{27^2}{144}\right)$$

$$P(163 \leq \overline{X} \leq 171) = P\left(\frac{163 - 167}{\frac{27}{\sqrt{144}}} \leq Z \leq \frac{171 - 167}{\frac{27}{\sqrt{144}}}\right) = 0.92$$

$$\Downarrow$$

Aumentando la numerosità del campione da 36 a 144, la probabilità richiesta è salita dal 63% al 92% circa.

Esercizio

La durata, in giorni, di alcuni apparecchi fatti funzionare ininterrottamente fino a che si rompono segue una distribuzione normale di media $\mu = 98$ e scarto quadratico medio $\sigma = 3.5$ ore. Selezionato un campione casuale di n= 10 apparecchi, determinare:

- La varianza della media campionaria
- la probabilità che la durata media di funzionamento sia compresa tra 95 e 101 ore;
- la probabilità che la durata media di funzionamento sia compresa tra 95 e 101 ore per un campione di ampiezza n= 20 apparecchi

Esercizio

Poichè la popolazione della Durata di funzionamento segue una distribuzione normale avente media $\mu = 98$, anche la media campionaria $\overline{X}$ segue una distribuzione normale avente media $\overline{X} = \mu = 98$.
La varianza invece sarà:

$$Var(\overline{X}) = \frac{\sigma^2}{n} = \frac{3.5^2}{10} = 1.225$$

Esercizio

la probabilità che la durata media di funzionamento sia compresa tra 95 e 101 ore

$$P(95 \leq \overline{X} \leq 101) = P\left(\frac{95-98}{\frac{3.5}{\sqrt{10}}} \leq \frac{\overline{X}-\mu}{\frac{\sigma}{n}} \leq \frac{101-98}{\frac{3.5}{\sqrt{10}}}\right)$$

$$P(-2.71 \leq Z \leq 2.71) = 0.4966 * 2 = 0.9932$$

Esercizio

la probabilità che la durata media di funzionamento sia compresa tra 95 e 101 ore per un campione di ampiezza n= 20 apparecchi
La media campionaria $\overline{X}$ segue una distribuzione normale avente media $\overline{X} = \mu = 98$ mentre la varianza invece sarà:

$$Var(\overline{X}) = \frac{\sigma^2}{n} = \frac{3.5^2}{20} = 0.6125$$

da cui si deduce che, al crescere della numerosità campionaria, la distribuzione media campionaria presenta minore variabilità dei dati intorno alla media.

$$P(95 \leq \overline{X} \leq 101) = P\left(\frac{95-98}{\frac{3.5}{\sqrt{20}}} \leq \frac{\overline{X}-\mu}{\frac{\sigma}{n}} \leq \frac{101-98}{\frac{3.5}{\sqrt{20}}}\right)$$

$$P(-3.83 \leq Z \leq 3.83) = 0.4999 * 2 = 0.9998$$

Esercizio

I perimetri toracici della popolazione maschile italiana, di età compresa tra i 18 e i 74 anni, si distribuiscono normalmente con media $\mu = 75cm$ e scarto quadratico medio $\sigma = 19$ cm.
Determinare la probabilità che il parametro toracico medio calcolato in un campione casuale di numerosità n =100 superi i 79.75 cm.

Esercizio

Sia $\overline{X}$ la distribuzione dei parametri toracici maschili, la probabilità richiesta si calcola nel modo seguente:

$$P(\overline{X} > 79.75) = P\left(\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} > \frac{79.75 - 75}{\frac{19}{\sqrt{100}}}\right) = P(z > 2.5) = 0.00621$$

Riepilogo

Abbiamo visto che, data una v.c. X con μ *e* σ^2 finite ma incognite, degli stimatori accettabili di questi parametri sono:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \textit{per } \mu$$

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 \quad \textit{per } \sigma^2, \textit{ se } \mu \textit{ e' nota}$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \quad \textit{per } \sigma^2, \textit{ se } \mu \textit{ non e' nota}$$

Dato che questi stimatori sono delle v.c. con distribuzione di probabilità dipendente da quella della v.c. di partenza, sorge il problema di derivare, in modo esatto o approssimato, la loro distribuzione in modo da potere inferire sulle relative proprietà statistiche.