

Formule e concetti principali

1 Introduzione alla statistica

Statistica: l'arte di apprendere dai dati

Statistica descrittiva: descrive e riassume i dati

Statistica inferenziale: trarre conclusioni dai dati

Popolazione: insieme di elementi di interesse

Campione: la parte della popolazione da cui si ottengono i dati

2 Descrivere insiemi di dati

Tabelle e grafici delle frequenze e delle frequenze relative

Istogrammi

Diagrammi ramo-foglia

Diagrammi di dispersione per dati a coppie

3 Uso delle statistiche per riassumere i dati

Media campionaria: $\bar{x} = (\sum_{i=1}^n x_i)/n$

Mediana campionaria: il valore intermedio

Varianza campionaria: $s^2 = \sum_{i=1}^n (x_i - \bar{x})^2 / (n - 1)$

Deviazione standard campionaria: $s = \sqrt{s^2}$

Uguaglianza algebrica: $\sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - n\bar{x}^2$

Regola empirica per gli insiemi di dati normali:

circa 68% dei dati sono compresi in $\bar{x} \pm s$

circa 95% dei dati sono compresi in $\bar{x} \pm 2s$

circa 99.7% dei dati sono compresi in $\bar{x} \pm 3s$

Coefficiente di correlazione campionaria: $r = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) / [(n - 1)s_x s_y]$

4 Probabilità

$0 \leq P(A) \leq 1$

$P(S) = 1$, dove S è l'insieme di tutti i valori possibili

$P(A \cup B) = P(A) + P(B)$, quando A e B sono disgiunti

Probabilità del complementare: $P(A^c) = 1 - P(A)$

Regola dell'addizione: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Probabilità condizionata: $P(B|A) = P(A \cap B)/P(A)$

Regola della moltiplicazione: $P(A \cap B) = P(A)P(B|A)$

Eventi indipendenti: $P(A \cap B) = P(A)P(B)$

5 Variabili aleatorie discrete

Valore atteso (o media): $E[X] = \sum_{i=1}^n x_i P\{X = x_i\}$

$$E[X + Y] = E[X] + E[Y]$$

Varianza: $\text{Var}(X) = E[(X - E[X])^2] = E[X^2] - (E[X])^2$

Deviazione standard: $\text{SD}(X) = \sqrt{\text{Var}(X)}$

$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$ se X e Y sono indipendenti:

Variabile aleatoria binomiale: $P\{X = i\} = \frac{n!}{i!(n-i)!} p^i (1-p)^{n-i}, i = 0, \dots, n$

$$E[X] = np \quad \text{Var}(X) = np(1-p)$$

6 Variabili aleatorie normali

Variabile aleatoria normale X : caratterizzata da $\mu = E[X], \sigma = \text{SD}(X)$

Variabile aleatoria normale standard Z : normale con $\mu = 0, \sigma = 1$

$$P\{|Z| > x\} = 2P\{Z > x\}, x > 0$$

$$P\{Z < -x\} = P\{Z > x\}$$

z_α è tale che $P\{Z > z_\alpha\} = \alpha$

Se X è normale, allora $Z = (X - \mu)/\sigma$ è una normale standard.

Proprietà additiva: se X e Y sono normali indipendenti, $X + Y$ è normale con media $\mu_x + \mu_y$, e varianza $\sigma_x^2 + \sigma_y^2$

7 La distribuzione campionaria delle statistiche

$X_1, \dots, X_n$ è un campione della popolazione: $E[X_i] = \mu, \text{Var}(X_i) = \sigma^2$

$$\begin{aligned} E[\bar{X}] &= \mu \\ \text{Var}(\bar{X}) &= \sigma^2/n \end{aligned}$$

Teorema centrale del limite: $\sum_{i=1}^n X_i$ per grandi valori di n , è approssimativamente normale con media $n\mu$ e deviazione standard $\sigma\sqrt{n}$; analogamente $\sqrt{n}(\bar{X} - \mu)/\sigma$ ha distribuzione approssimativamente normale standard.

Approssimazione normale della binomiale: se $np \geq 5, n(1-p) \geq 5$ allora $[\text{Bin}(n, p) - np]/\sqrt{np(1-p)}$ è approssimativamente normale standard.

8 Stima

$\bar{X}$ è lo stimatore della media della popolazione μ .

$\hat{p}$, la proporzione del campione che ha una certa proprietà, stima p , la proporzione della popolazione che ha la proprietà.

S^2 stima σ^2 , e S stima σ .

Intervallo di confidenza al $100(1 - \alpha)\%$ per μ :

dati normali o n grande, σ nota: $\bar{X} \pm z_{\alpha/2} \sigma/\sqrt{n}$

dati normali, σ incognita: $\bar{X} \pm t_{n-1, \alpha/2} S/\sqrt{n}$

Intervallo di confidenza al $100(1 - \alpha)\%$ per p : $\hat{p} \pm z_{\alpha/2} \sqrt{\hat{p}(1 - \hat{p})/n}$

9 Verifica delle ipotesi statistiche

H_0 = ipotesi nulla, ipotesi sotto verifica

Livello di significatività α : la (più grande possibile) probabilità di rifiutare H_0 quando essa è vera
valore- p : il più piccolo livello di significatività a cui H_0 viene rifiutata

Verifica di ipotesi riguardanti la media μ di una popolazione

Assunzione: la distribuzione è normale, o n è grande.

H_0	H_1	Statistica del test ST	Test di livello α	Valore-p se $ST = v$
$\mu = \mu_0$	$\mu \neq \mu_0$	$\frac{\sqrt{n}(\bar{X} - \mu_0)}{\sigma}^\dagger$	Rifiutare H_0 se $ ST \geq z_{\alpha/2}$	$2P\{z \geq v \}$
$\mu \leq \mu_0$	$\mu > \mu_0$	$\frac{\sqrt{n}(\bar{X} - \mu_0)}{\sigma}^\dagger$	Rifiutare H_0 se $ST \geq z_\alpha$	$P\{Z \geq v\}$
$\mu = \mu_0$	$\mu \neq \mu_0$	$\frac{\sqrt{n}(\bar{X} - \mu_0)}{\sigma}$	Rifiutare H_0 se $ ST \geq t_{n-1, \alpha/2}$	$2P\{T_{n-1} \geq v \}$
$\mu \leq \mu_0$	$\mu > \mu_0$	$\frac{\sqrt{n}(\bar{X} - \mu_0)}{\sigma}$	Rifiutare H_0 se $ST \geq t_{n-1, \alpha}$	$P\{T_{n-1} \geq v\}$

† Assunzione: σ è nota.

Nota: per verificare $H_0: \mu \geq \mu_0$, moltiplicare i dati per -1 e utilizza il test già noto.

10 Verifica di ipotesi su due popolazioni

Verifica delle medie di due popolazioni quando i campioni sono indipendenti

Il campione X di numerosità n e il campione Y di numerosità m sono indipendenti.

H_0	H_1	Statistica del test ST	Assunzione	Test di livello α	Valore-p se $ST = v$
$\mu_x = \mu_y$	$\mu_x \neq \mu_y$	$\frac{\bar{X} - \bar{Y}}{\sqrt{S_x^2/n + S_y^2/m}}$	n, m grandi	Rifiutare se $ ST \geq z_{\alpha/2}$	$2P\{Z \geq v \}$
$\mu_x \leq \mu_y$	$\mu_x > \mu_y$	$\frac{\bar{X} - \bar{Y}}{\sqrt{S_x^2/n + S_y^2/m}}$	n, m grandi	Rifiutare se $ST \geq z_\alpha$	$P\{Z \geq v\}$
$\mu_x = \mu_y$	$\mu_x \neq \mu_y$	$\frac{\bar{X} - \bar{Y}}{\sqrt{S_p^2(1/n + 1/m)}}$	Popolazioni normali $\sigma_x = \sigma_y$	Rifiutare se $ST \geq t_{n+m-2, \alpha/2}$	$2P\{T_{n+m-2} \geq v \}$
$\mu_x \leq \mu_y$	$\mu_x > \mu_y$	$\frac{\bar{X} - \bar{Y}}{\sqrt{S_p^2(1/n + 1/m)}}$	Popolazioni normali $\sigma_x = \sigma_y$	Rifiutare se $ST \geq t_{n+m-2, \alpha}$	$P\{T_{n+m-2} \geq v\}$

$$S_p^2 = \frac{n-1}{n+m-2} S_x^2 + \frac{m-1}{n+m-2} S_y^2 = \text{stimatore combinato del valore comune di } \sigma_x^2 = \sigma_y^2$$

Verifica di ipotesi riguardanti p

(la proporzione di una grande popolazione che ha una certa caratteristica)

X è il numero di membri di una popolazione in un campione di numerosità n che hanno la caratteristica. B è una variabile aleatoria binomiale con parametri n e p_0 .

H_0	H_1	Statistica del test ST	Valore-p se $ST = x$
$p \leq p_0$	$p > p_0$	X	$P\{B \geq x\}$
$p = p_0$	$p \neq p_0$	X	$2 \min \{P\{B \leq x\}, P\{B \geq x\}\}$

Test di uguaglianza di proporzioni di popolazione

p_1 e p_2 sono le proporzioni dei membri di due popolazioni che hanno una certa caratteristica di interesse. Un campione casuale indipendente di numerosità n_1 viene estratto dalla prima popolazione e un campione casuale indipendente di numerosità n_2 viene estratto dalla seconda. $\hat{p}_1$ e $\hat{p}_2$ sono le proporzioni dei membri dei due campioni che hanno la caratteristica e $\hat{p}$ è la proporzione dei campioni combinati che hanno la caratteristica.

H_0	H_1	Statistica del test ST	Test di livello α	Valore- p se $ST = v$
$p_1 = p_2$	$p_1 \neq p_2$	$\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{(1/n_1 + 1/n_2)\hat{p}(1 - \hat{p})}}$	Rifiuta H_0 se $ ST \geq z_{\alpha/2}$	$2P\{Z \geq v \}$
$p_1 \leq p_2$	$p_1 > p_2$	$\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{(1/n_1 + 1/n_2)\hat{p}(1 - \hat{p})}}$	Rifiuta H_0 se $ST \geq z$	$P\{Z \geq v\}$

11 Analisi della varianza

Tabella ANOVA a un fattore

Le variabili $\bar{X}_i$ e $S_i^2, i = 1, \dots, m$, sono rispettivamente le medie e le varianze campionarie di campioni indipendenti di numerosità n per popolazioni con medie μ_i e la stessa varianza σ^2 .

Sorgente dello stimatore	Stimatore di σ^2	Valore della statistica del test
Tra campioni	$n\bar{S}^2 = \frac{n \sum_{i=1}^m (\bar{X}_i - \bar{\bar{X}})^2}{(m - 1)}$	$ST = \frac{n\bar{S}^2}{\left(\sum_{i=1}^m S_i^2\right)/m}$
All'interno di ciascun campione	$\left(\sum_{i=1}^m S_i^2\right)/m$	

Test di livello α di H_0 : tutti i valori μ_i sono uguali

Rifiutare H_0 se $ST \geq F_{m-1, m(n-1), \alpha}$

Non rifiutare H_0 altrimenti

Se $ST = v$ allora

$$\text{valore-}p = P\{F_{m-1, m(n-1)} \geq v\}$$

dove $F_{m-1, m(n-1)}$ è una variabile aleatoria F con $m - 1$ gradi di libertà al numeratore e $m(n - 1)$ al denominatore.

Modello ANOVA a due fattori: per $i = 1, \dots, m, j = 1, \dots, n$

$$E[X_{ij}] = \mu + \alpha_i + \beta_j$$

$$\sum_{i=1}^m \alpha_i = \sum_{j=1}^n \beta_j = 0$$

il valore μ si chiama media generale, α_i è la deviazione dalla media generale dovuta alla riga i , e β_j è la deviazione dalla media generale dovuta alla colonna j . I loro stimatori sono

$$\hat{\mu} = X_{..} \quad \hat{\alpha}_i = X_{i.} - X_{..} \quad \hat{\beta}_j = X_{.j} - X_{..}$$

Tabella ANOVA a due fattori

	Somma dei quadrati	Gradi di libertà
Riga	$SS_r = n \sum_{i=1}^m (X_{i.} - X_{..})^2$	$m - 1$
Colonna	$SS_c = m \sum_{j=1}^n (X_{.j} - X_{..})^2$	$n - 1$
Errore	$SS_e = \sum_{i=1}^m \sum_{j=1}^n (X_{ij} - X_{i.} - X_{.j} + X_{..})^2$	$N = (n - 1)(m - 1)$

Ipotesi nulla	Statistica del test	Test di livello α	Valore- p se $ST = \nu$
Nessun effetto di riga (tutte le $\alpha_i = 0$)	$\frac{SS_r/(m - 1)}{SS_e/N}$	Rifiutare se $ST \geq F_{m-1, N, \alpha}$	$P\{F_{m-1, N} \geq \nu\}$
Nessun effetto di colonna (tutte le $\beta_i = 0$)	$\frac{SS_c/(n - 1)}{SS_e/N}$	Rifiutare se $ST \geq F_{n-1, N, \alpha}$	$P\{F_{n-1, N} \geq \nu\}$

12 Regressione lineare

Modello di regressione lineare semplice: $Y = \alpha + \beta x + e$

Stimatori dei minimi quadrati: $\hat{\beta} = S_{xy}/S_{xx}$, $\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{x}$

$$S_{xy} = \sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y}) = \sum_{i=1}^n x_i Y_i - n\bar{x}\bar{Y}$$

$$S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - n\bar{x}^2$$

Retta di regressione stimata: $y = \hat{\alpha} + \hat{\beta}x$

L'errore e è una variabile aleatoria normale con media 0 e varianza σ^2 . Lo stimatore di σ^2 è $SS_R/(n - 2)$,

$$SS_R = \sum_i (Y_i - \hat{\alpha} - \hat{\beta} x_i)^2 = (S_{xx}S_{yy} - S_{xy}^2)/S_{xx}$$

Per verificare $H_0: \beta = 0$. Utilizza $ST = \sqrt{(n - 2)S_{xx}/SS_R} \hat{\beta}$

Un test di livello γ consiste nel rifiutare H_0 se $|ST| \geq t_{n-2, \gamma/2}$.

Se $ST = \nu$, valore- $p = 2P\{T_{n-2} \geq \nu\}$

100(1 - γ) intervallo di predizione per una risposta al predittore x_0

$$\hat{\alpha} + \hat{\beta}x_0 \pm t_{n-2, \gamma/2} \sqrt{(1 + 1/n + (x_0 - \bar{x})^2/S_{xx})SS_R/(n - 2)} \hat{\alpha}$$

Coefficiente di determinazione: $R^2 = 1 - SS_R/S_{yy}$ rappresenta la proporzione della variazione delle variabili risposta che è giustificata dai valori dei predittori. La sua radice quadrata è il valore assoluto del coefficiente di correlazione campionaria.

Modello di regressione lineare multipla:

$$Y = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k + e$$

13 Test chi-quadrato di adattamento

P_i è la proporzione della popolazione con valore i , $i = 1, \dots, k$.

Per verificare $H_0: P_i = p_i$, $i = 1, \dots, k$, consideriamo un campione di numerosità n .

Sia N_i be il numero di elementi del campione che hanno il valore i , $e_i = np_i$, $ST = \sum_{i=1}^k (N_i - e_i)^2 / e_i$. Il test di livello α rifiuta H_0 se $ST \geq \chi_{k-1, \alpha}^2$.

Se $ST = v$, allora valore- $p = P[\chi_{k-1}^2 \geq v]$.

Supponiamo che ciascun membro di una popolazione abbia una caratteristica X e una caratteristica Y . Ipotizziamo r valori possibili della caratteristica X e s valori possibili della caratteristica Y . Per verificare l'ipotesi che le caratteristiche di un membro della popolazione scelto a caso siano indipendenti, consideriamo un campione casuale di numerosità n .

N_{ij} = numero degli elementi del campione che hanno la caratteristica X uguale a i e la caratteristica Y uguale a j

N_i = numero degli elementi del campione che hanno la caratteristica X uguale a i

M_j = numero degli elementi del campione che hanno la caratteristica Y uguale a j

$\hat{e}_{ij} = N_i M_j / n$

Se $\sum_i \sum_j (N_{ij} - \hat{e}_{ij})^2 / \hat{e}_{ij} \geq \chi_{(r-1)(s-1), \alpha}^2$ allora l'ipotesi di indipendenza è rifiutata per un livello di significatività α .

14 Test non parametrici

Sia η = mediana della popolazione. Il test dei segni di

$$H_0: \eta = m \quad \text{contro} \quad H_1: \eta \neq m$$

richiede un campione di numerosità n . Se i elementi sono minori di m , allora

$$\text{valore-}p = 2 \min (P\{N \leq i\}, P\{N \geq i\})$$

dove N è una variabile aleatoria binomiale $(n, 1/2)$.

Il test dei ranghi con segno è utilizzato per verificare l'ipotesi che una distribuzione di popolazione sia simmetrica rispetto a 0. Esso classifica i dati in base al valore assoluto. ST è la somma dei ranghi dei valori negativi

$$\text{valore-}p = 2 \min (P\{ST \leq t\}, P\{ST \geq t\})$$

ST è approssimativamente normale con media $n(n+1)/4$ e varianza $n(n+1)(2n+1)/24$.

Per verificare l'uguaglianza di due distribuzioni di popolazione, estraiamo due campioni casuali di numerosità n e m e mettiamo in ordine gli $n+m$ valori. Il test della somma dei ranghi utilizza ST = somma dei ranghi del primo campione. Il test rifiuta H_0 se ST è molto grande o molto piccola. Se $ST = t$, allora

$$\text{valore-}p = 2 \min (P\{ST \leq t\}, P\{ST \geq t\})$$

ST è approssimativamente normale con media $n(n+m+1)/2$ e varianza $nm(n+m+1)/12$.

Per verificare l'ipotesi che una successione di cifre 0 e 1 sia casuale, utilizziamo il test delle sequenze contando R , il numero di sequenze. Rifiutiamo la casualità quando R è troppo piccolo o troppo grande per essere giustificabile dal caso. Utilizziamo il risultato che quando H_0 è vera, R è approssimativamente normale con media $1 + 2nm/(n+m)$ e varianza

$$\frac{2nm(2nm - n - m)}{(n+m)^2(n+m-1)}$$

15 Controllo di qualità

Limiti di carta di controllo $\mu \pm 3\sigma/\sqrt{n}$ n = numerosità del sottogruppo

Area sottesa dalla curva normale standard a sinistra di x

x	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5733
.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998