Statistica e analisi dati

Nozioni base di probabilità

$P(A)+P(\neg A)=1$
$P(A\cup B)+P(A\cap B)=P(A)+P(B)$
$P(A\cap B)=P(A)\cdot P(B|A)$
$P(A\cap B)=P(A)\cdot P(B)$ se A e B sono indipendenti.
$P(B)=\sum_{i=1}^{n}P(A_i)P(B|A_i)$

Teorema di Bayes

$P(A|B)=\frac{P(A)P(B|A)}{P(B)}=\frac{P(A\cap B)}{P(B)}$

Nozioni base di statistica

Momento sul discreto: $\langle x^r\rangle=\frac{1}{n}\sum_{i=1}^{n}x_i^r$
Momento sul continuo: $\langle x^r\rangle=\int_{-\infty}^{+\infty}x^r f(x)\,dx$
Media/valore atteso: $\mu=\bar{x}=\langle x\rangle$
Mediana: $F(x)=\frac{1}{2}$
Primo quartile: $F(x)=\frac{1}{4}$
Terzo quartile: $F(x)=\frac{3}{4}$
Intervallo interquartile: $\Delta=x_{(3°\text{ quartile})}-x_{(1°\text{ quartile})}$
Varianza: $\sigma^2=\left\langle(x-\bar{x})^2\right\rangle=\langle x^2\rangle-\langle x\rangle^2$
Deviazione standard: $\sigma=\sqrt{\sigma^2}$
Momento standard: $\mu_r=\frac{\langle(x-\bar{x})^r\rangle}{\sigma^r}$
Skweness: μ_3
Curtosi: μ_4
Funzione di fallibilità: $F(x)=\int_{-\infty}^{x}f(t)dt$
Funzione di sopravvivenza: $S(x)=1-F(x)$

Teorema di Chebyshev

In un intervallo entro due volte la deviazione standard dalla media, è contenuto almento il 75% della probabilità.

Distribuzioni

Distribuzione uniforme

Funzione di densità: $f(t)=\begin{cases}\frac{1}{b-a} & \text{se } a\le t\le b\\ 0 & \text{altrimenti}\end{cases}$

Funzione cumulativa: $F(t)=\begin{cases}0 & \text{se } t<a\\ \frac{x-a}{b-a} & \text{se } a\le t\le b\\ 1 & \text{se } t>b\end{cases}$

Media/valore atteso: $\frac{a+b}{2}$
Mediana: $\frac{a+b}{2}$
Varianza: $\frac{(b-a)^2}{12}$

Distribuzione geometrica

La distribuzione geometrica esprime la probabilità che occorra attendere esattamente i tentativi per avere il primo successo. La distribuzione geometrica è **senza memoria**.

Funzione di densità: $\mathcal{G}(i\,|\,p)=pq^{i-1}$
Funzione cumulativa: $1-q^i$
Media/valore atteso: $\frac{1}{p}$
Moda: 1
Varianza: $\frac{q}{p^2}$

Distribuzione binomiale

La distribuzione binomiale esprime la probabilità di avere esattamente k successi su n tentativi.

Funzione di densità: $\mathcal{B}(k\,|\,p,n)=\binom{n}{k}p^k q^{n-k}$
Media/valore atteso: np
Mediana: $\lfloor np\rfloor$ o $\lceil np\rceil$
Moda: $\lfloor(n+1)p\rfloor$ o $\lceil(n+1)p\rceil-1$
Varianza: npq

Distribuzione esponenziale

La distribuzione esponenziale esprime la probabilità di attendere esattamente un tempo t per avere il primo evento.
La distribuzione esponenziale è **senza memoria**.

Funzione di densità: $f(t\,|\,\lambda)=\lambda e^{-\lambda t}$
Funzione cumulativa: $1-e^{-\lambda t}$
Media/valore atteso: $\frac{1}{\lambda}$
Mediana: $\frac{\ln 2}{\lambda}$
Moda: 0
Varianza: $\frac{1}{\lambda^2}$

Distribuzione di Poisson

La distribuzione di Poisson esprime la probabilità di avere esattamente k eventi in un intervallo di tempo quando la media di eventi è μ.
La distribuzione di Poisson viene anche usata per approssimare la distribuzione binomiale quando n è molto grande e p molto piccolo.
Data una distribuzione esponenziale, la relativa distribuzione di conteggio è una distribuzione di Poisson dove $\mu=\lambda\Delta t$.

Funzione di densità: $\mathcal{P}(k\,|\,\mu=np)=\frac{\mu^k}{k!}e^{-\mu}$
Media/valore atteso: μ
Moda: $\lceil\mu\rceil-1$ e $\lfloor\mu\rfloor$
Varianza: μ
Merge: ovrapponendo due processi Poissoniani con rate λ_1 e λ_2, ottengo un processo Poissoniano di rate λ.
Split: dato un processo Poissoniano di rate λ, estraendo ogni evento con probabilità p, ottengo due processi Poissoniani di rate $p\lambda$ e $(1-p)\lambda$.

Distribuzione normale (Gaussiana)

La distribuzione di Poisson viene usata per approssimare la distribuzione binomiale quando n è molto grande e p è “lontato” da 0 e 1.

Funzione di densità: $\mathcal{N}(x\,|\,\mu,\sigma)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}$
Media/valore atteso: μ
Mediana: μ
Moda: μ
Varianza: σ^2
Standardizzazione: $\mathcal{N}(x\,|\,\mu,\sigma)=\mathcal{N}(z\,|\,0,1)$, per $z=\frac{x-\mu}{\sigma}$
Legge tre sigma:

- $P(\mu-1\sigma\le X\le\mu+1\sigma)\approx 68{,}27\%$
- $P(\mu-2\sigma\le X\le\mu+2\sigma)\approx 95{,}45\%$
- $P(\mu-3\sigma\le X\le\mu+3\sigma)\approx 99{,}73\%$

Distribuzione ipergeometrica

La distribuzione ipergeometrica esprime la probabilità di estrarre senza reinserimento g palline vincenti su n estratte da un'urna contenente G palline vincenti e B palline perdenti.

Funzione di densità: $\mathcal{H}(g\,|\,n,G,B)=\frac{\binom{G}{g}\binom{B}{n-g}}{\binom{G+B}{n}}$

Somma di variabili aleatorie

Media: $\mu_Z=\mu_X+\mu_Y$
Varianza: $\sigma_Z^2=\sigma_X^2+\sigma_Y^2$ se X e Y sono indipendenti

Distribuzioni campionarie

Minimo campionario: $f_{\min}(t)=nf(t)\,(S\,(t))^{n-1}$
Massimo campionario: $f_{\max}(t)=nf(t)\,(F\,(t))^{n-1}$
Media campionaria: una distribuzione di media μ e varianza $\frac{\sigma^2}{n}$

Teorema del limite centrale

Sommando variabili aleatorie indipendenti con distribuzioni qualsiasi, purché dotate di varianza finita, ottengo, nel limite, una variabile Gaussiana: $f_{\text{avg}}(t)=\mathcal{N}\left(t\,|\,\mu,\frac{\sigma^2}{n}\right)$

Correlazione

Codevianza: $\text{cod}(x,y)=\sum_{i=1}^{n}(x_i-\mu_x)(y_i-\mu_y)$
Covarianza: $\text{cov}(x,y)=\frac{\text{cod}(x,y)}{n}$
Coefficiente di Pearson: $\rho_{x,y}=\frac{\text{cov}(x,y)}{\sigma_x\sigma_y}$

- $\rho<0$: correlazione negativa;
- $\rho=0$: nessuna correlazione;
- $\rho>0$: correlazione positiva;
- $0\le|\rho|<0{,}3$: correlazione debole;
- $0{,}3\le|\rho|<0{,}7$: correlazione moderata;
- $0{,}7\le|\rho|<1$: correlazione forte;
- $|\rho|=1$: correlazione perfetta;

Coefficiente di Spearman: $r_s=\rho_{\text{R}(X),\,\text{R}(Y)}=\frac{\text{cov}(\text{R}(X),\,\text{R}(Y))}{\sigma_{\text{R}(X)}\sigma_{\text{R}(Y)}}$
Dove $R(x)$ è il rango di x, ovvero la posizione di x all'intero di X.

Stima della media

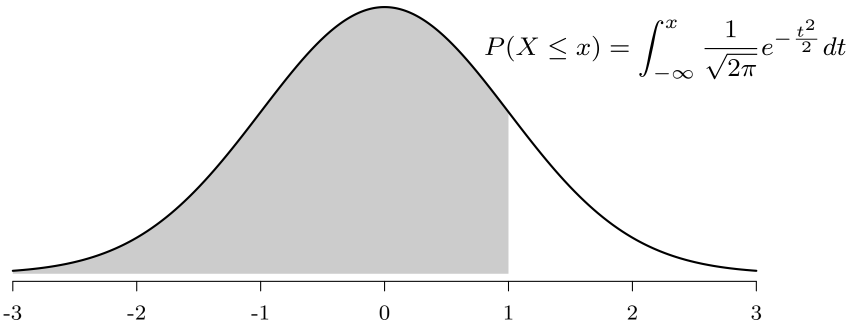
Dato un campione di n eventi indipendenti da una popolazione con varianza finita σ^2 e media empirica m, la media “vera” è distribuita secondo una distribuzione gaussiana con media m e deviazione standard $\frac{\sigma}{\sqrt{n}}$.

Con varianza non nota si usa la varianza empirica: $s^2=\frac{\sum(x_i-m)^2}{n-1}$
Media stimata: $\mu_{(\text{stima})}=m\pm\frac{\sigma}{\sqrt{n}}$
Intervallo di confidenza del 68%: $[m-\ \frac{\sigma}{\sqrt{n}}\,;\,m+\ \frac{\sigma}{\sqrt{n}}]$
Intervallo di confidenza del 95%: $[m-2\frac{\sigma}{\sqrt{n}}\,;\,m+2\frac{\sigma}{\sqrt{n}}]$
Intervallo di confidenza del 99,7%: $[m-3\frac{\sigma}{\sqrt{n}}\,;\,m+3\frac{\sigma}{\sqrt{n}}]$
Stima della probabilità in una dist. bernulliana: $p_{(\text{stima})}=\frac{k}{n}\pm\frac{\sqrt{k}}{n}$

Basato sul corso *Statistica e analisi dei dati*
A.A. 2021/2022 del docente Gianini Gabriele.

Tavola della distribuzione normale standardizzata

$\mathcal{N}(x \mid \mu = 0, \sigma = 1)$



$$P(X \leq x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} \, dt$$

	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990