Statistica

Il data science si riduce all'applicazione di metodi statistici, questi metodi servono a collezionare dati, analizzarli e descriverli.

Sarà possibile fare inferenze, cioè cercare di trarre delle conclusioni su una popolazione avendo un campione significativo.

Data types

Le variabili possono essere:

- Categoriche, rappresentano un gruppo di oggetti, risposte si/no, divise a loro volta in:
- Nominali, i gruppi non possono essere messi in una scala.
- Ordinali, i gruppi possono essere messi in una scala.
- Numeriche, divise in:
- Discrete
- Continue

Experimental probability

- Trial: osservazione di un evento e registro il risultato.
- Esperimento: un insieme di trial.
- Sample space: tutti i possibili risultati che posso ottenere.
- Probabilità: rapporto tra l'evento che mi interessa e tutti i risultati possibili.
- Valore atteso: valore medio che mi aspetto quando ripeto l'esperimento molte volte. Si ottiene sommando i prodotti fra probabilità di un certo valore per il valore stesso.

Per esempio avendo due dadi da 6 facce il valore medio che uscirà sarà 7 perchè è quello con più combinazioni possibili.

SALTIAMO DA SLIDE 10 A 15

Distribuzione di probabilità

Descrive tutti i valori assunti dalla variabile e ad ciascun valore associa il conteggio o la probabilità. La distribuzione viene descritta tramite:

- media [μ]: somma di tutti i valori / numero di valori,

- varianza: misura di quanto la distribuzione è "appiattita"/allargata. La somma di tutte le distanze fra individuo e media, ciascuna al quadrato / numero di individui,
- deviazione standard [α]: radice quadrata della varianza,

Misure di tendenza centrale

Ecco alcune misure di tendenza centrale:

- media: come prima,
- mediana: numero nel mezzo di un dataset ordinato,
- moda: valore più frequente.

Misure di asimmetria

Qui la misura principale è la skewness, cioè se i dati sono concentrati da un lato preciso.

Se media > mediana abbiamo una skew positiva quindi i dati si concentrano verso il lato SX con coda a DX (valori estremi che spostano la media). Invece è negativa se media < mediana e si ha una coda verso SX.

Misure di variabilità

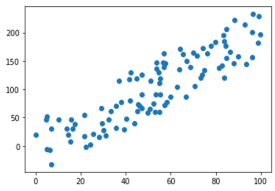
Ci sono la varianza e la deviazione standard.

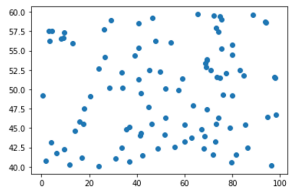
Misura di correlazione

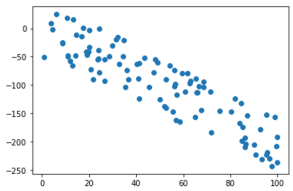
Misura di quanto due variabili siano correlate tra loro, sia ha:

- Covarianza
- Correlazione lineare: esiste il metodo su pandas, valore compreso fra -1 e 1, più è vicino a 1 più c'è correlazione (si seguono), più è vicino a -1 c'è correlazione opposta (prendono 'strade diverse') e con 0 non c'è tipo di correlazione.

In un scatter plot sono rappresentate in questo modo:







Distribuzioni discrete e continue

Nelle discrete ogni risultato univoco è assegnata una probabilità, nelle continue ogni valore può assumere tutti i valori in un certo intervallo, quindi non possiamo

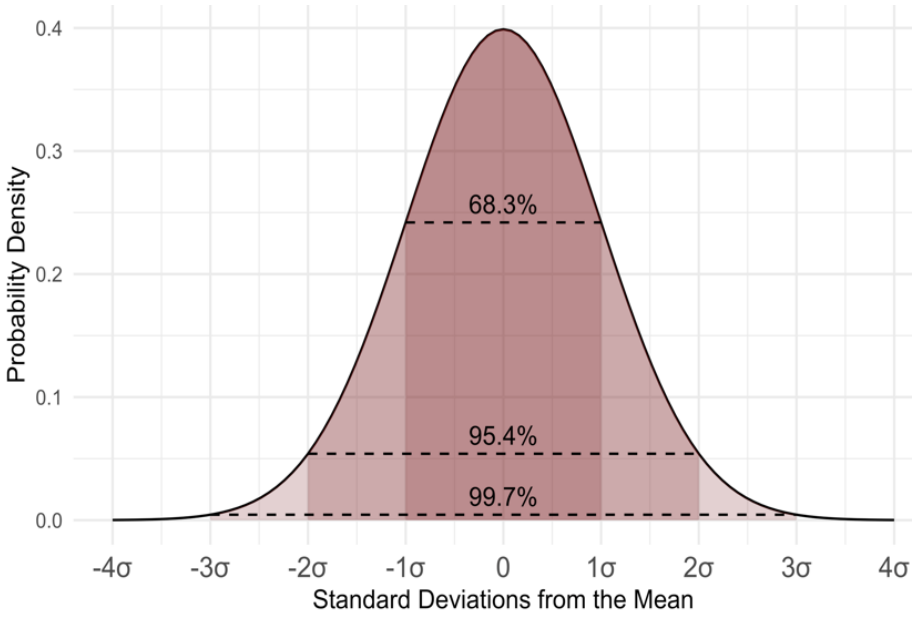
associare ad ogni valore preciso una percentuale ma la assoceremo ad un range di valori.

Alcuni esempi di discreta sono:

- Distribuzione uniforme: ogni variabile ha la stessa probabilità di assumere qualsiasi valore all'interno di un intervallo specificato, es: lancio di un dado, ogni faccia ha la stessa % di uscita.
- Distribuzione binomiale: distribuzione del possibile numero di esiti positivi in un dato numero di prove in ciascuna delle quali esiste la stessa probabilità di successo.

Alcuni esempi di continue:

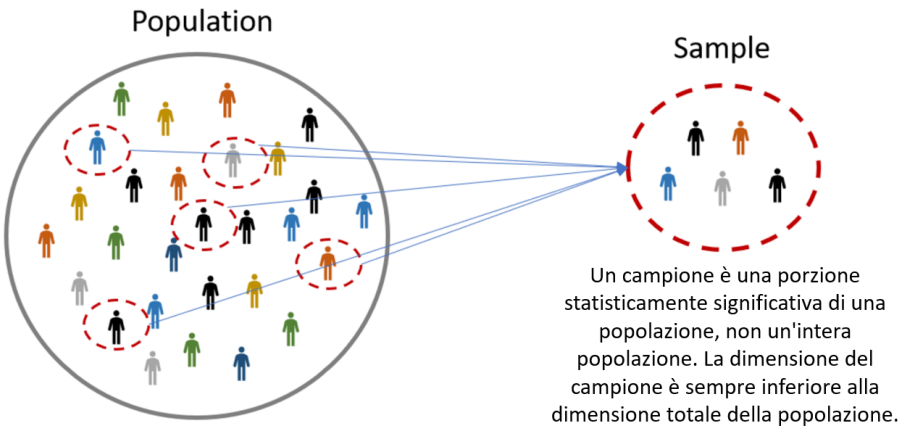
- Normal distribution: o gaussiana, è simmetrico rispetto alla media, mostrando che i dati vicini alla media sono più frequenti rispetto ai dati lontani dalla media. Sostanzialmente ha la seguente proprietà: se partendo dalla media tolgo o aumento una deviazione standard avrò un intervallo dove si avrà il 68,3 % della popolazione, se mi sposto di due deviazioni avrò il 95.4% della popolazione.



Di sotto c'è la tabella che rappresenta di quanto dobbiamo aggiungere o togliere per avere la percentuale di popolazione:

z	0	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

Popolazione e campione



Teorema del limite centrale

Prendendo un insieme di campioni random da una popolazione più grande e facendo la media otterremo un valore che va bene anche per l'intera popolazione, questo è vero grazie a questo teorema.

Prendendo più campioni e facendo sempre le medie e mettendole tutte in un grafico queste si distribuiranno tramite una gaussiana/distribuzione normale e non importa com'era la distribuzione di partenza.

Errore standard

Se nel teorema precedente prendiamo campioni sempre più piccoli la media risultante sarà affetta da un errore sempre più grande, al contrario più il campione è grande più l'errore diminuisce.

L'errore standard segue la formula:

deviazione standard / radice del numero di campioni

Intervalli e livelli di confidenza

Come detto prima se dalla media ci spostiamo di + o - due (1.960 per l'esattezza) valori dalla deviazione standard avremo il 95% circa della popolazione, quindi

abbiamo la sicurezza che ad ogni calcolo al 95% la media cada la dentro, ma è possibile che un campione cadi fuori con una percentuale del 5%

L'intervallo di confidenza è un range, che va dalla media - l'errore alla media + l'errore, con la media nel mezzo, all'interno del quale si stima che possa trovarsi il vero valore di un parametro di interesse. Il livello invece è la probabilità, che la media calcolata tramite campione ha, di cadere lì dentro.

Per trovare l'intervallo c'è la formula:

$$media \pm z * errore\ standard$$

dove la z la prendiamo dalla tabella in modo da ottenere la percentuale voluta che il valore cada dentro l'intervallo.

Quindi usando la formula otteniamo due valori (uno calcolando con il + l'altro con il -) che sono gli estremi dell'intervallo che ha la percentuale, data da z, di quanto un valore possa cadere lì dentro.

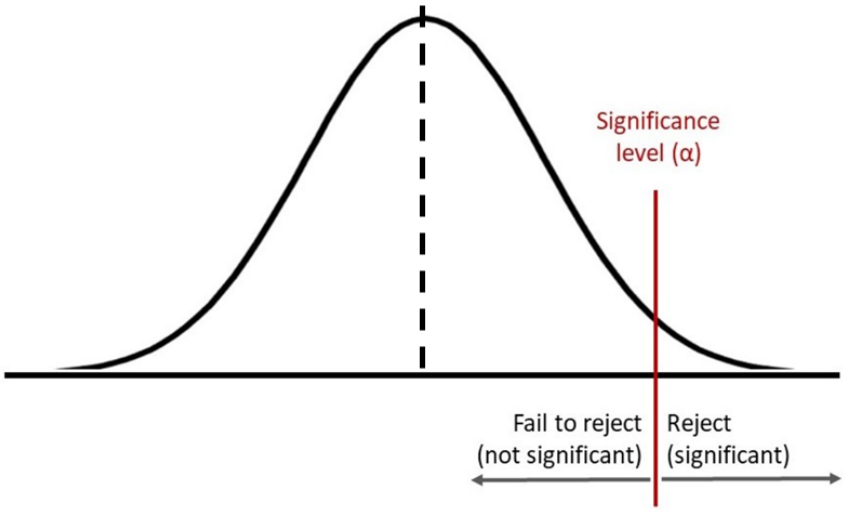
Test di ipotesi

Serve per verificare che un dato ottenuto da una popolazione sia effettivamente significativo.

I passaggi sono:

- imporre l'ipotesi nulla: il caso base, se per esempio vogliamo valutare se i tiri con i dadi che abbiamo fatto siano veri l'ipotesi nulla è dire che il dado sia regolare e che quindi tutte le facce escano con la stessa frequenza;
- settare un livello di significabilità α: per esempio voglio essere sicuro al 95% quindi alfa sarà 95:
- calcolo la probabilità: il p-value, cioè quanto è raro che il risultato ottenuto possa avvenire.

Per concludere se il p-value $< \alpha$ allora posso rifiutare/rigettare l'ipotesi nulla.



I test sono diversi e si differenziano in aspetti come il numero di campioni, e sono:

- T-test: confrontare due gruppi/categorie di variabili numeriche con un campione di piccole dimensioni
- Z-test: confrontare due gruppi/categorie di variabili numeriche con un campione di grandi dimensioni
- ANOVA test: confrontare la differenza tra due o più gruppi/categorie di variabili numeriche
- Chi-Squared test: esaminare la relazione tra due variabili categoriche
- Correlation test: esaminare la relazione tra due variabili numeriche

Regressione lineare

Fa parte del machine learning supervisionato, quindi forniamo l'input e label cioè la descrizione dell'input.

In questo caso abbiamo variabili indipendenti e fornisco l'output atteso in modo da allenare il modello.

La regressione lineare è un modello che cerca di trovare la relazione fra più variabili indipendenti e una variabile dipendente. Nel dettaglio il modello va a trovare la retta che descrive meglio la relazione fra i valori, quindi il coefficiente della retta che si avvicina di più ai valori sullo scatter plot.

Quindi avremo una funzione:

$$y = b_0 + b_1 \cdot x$$

Dove:

- y: variabile dipendente
- x: variabili indipendente
- b_0: intercetta, l'alzata all'origine cioè dove l'x è 0 l'y assume il valore di 0
- b_1: coefficiente angolare, la pendenza della retta e dice la variazione delle x al variare delle y.

Per individuare questa retta è definita una Cost function, che dà la somma delle misure fra la distanza dei punti e la retta (fra y e y) al quadrato diviso poi il numero di punti.

L'obiettivo del modello sarà poi quello di trovare i coefficienti in modo che la cost function sia la più piccola possibile.

Esiste l'R quadro che serve per capire la bontà del modello, per farlo calcoliamo la distanza fra i punti e la retta tracciata nel punto medio di y, perché se i punti sono distanti tra loro senza correlazione la retta creata dal modello non sarà mai buona come quella della media; in sostanza:

Se la retta è buona R^2 è uguale a 1

Time series

Nell'analizzare le serie storiche possiamo farlo in due modi: con scomposizione o modellistica (usato da noi).

Una serie storica è stazionaria se non c'è correlazione con il tempo e se oscilla tra un medesimo valore, inoltre non ci devono essere trend e stagionalità.

Nella funzione di autocorrelazione globale, detta anche ACF, i picchi determinano una correlazione fra ogni variabile della serie e k valori indietro.

Per esempio un picco su 1 indica che c'è correlazione tra ogni valore della serie e il valore precedente, se c'è sul 2 indica che c'è correlazione tra ogni valore della serie e il valore che ricorre due valori precedenti.