Statistica e Analisi dei Dati

June 29, 2020

S. Re, Statistica e analisi dei dati

1 Statistica descrittiva

1.1 Descrizione di dati

Popolazione: insieme di tutti gli elementi che ci interessano

Campione: sottoinsieme (rappresentativo) della popolazione che viene studiato

- Campione casuale semplice se i membri sono i scelti in modo tale che tutte le possibili scelte dei k membri siano equiprobabili
- Campione casuale stratificato se sono necessarie più informazioni iniziali

Frequenza assoluta (f): numero di occorrenza di un dato valore in un esperimento

Frequenza relativa: $\frac{f}{n}$

Tabelle e grafici: pochi valori distinti - Grafico a bastoncini - Grafico poligonale - Grafico a barre

Istogrammi: tanti dati, li suddivido in range distinti - Istogramma delle frequenze - Istograma delle frequenze relative

1.2 Riassumere i dati

Media campionaria: $\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$

- Traslazione: per i = 1, ..., n abbiamo

  $y_i = x_i + c$

  $\overline{y} = \overline{x} + c$

- Dilatazione: per i = 1, ..., n abbiamo

  $y_i = cx_i$

  $\overline{y} = c\overline{x}$

Media con frequenza: abbiamo k valori $x_1, x_2, ..., x_k$ con le relative frequenze $f_1, f_2, ..., f_k$ e numero di osservazioni totali pari a $n = \sum_{i=1}^{k} f_i$. Con queste premesse, posso calcolare la media campionaria come:

$\overline{x} = \frac{x_1+...+x_1+x_2+...+x_2+...x_k+...+x_K}{n} = \frac{f_1x_1+f_2x_2+...+f_kx_k}{n} = \frac{f_1}{n}x_1 + \frac{f_2}{n}x_2 + ...\frac{f_k}{n}x_k$

Ma le varie divisioni $\frac{f_i}{n}$ rappresentano la frequenza relativa, quindi possiamo vedere il tutto come:

$\overline{x} = w_1x_1 + w_2x_2 + ... + w_kx_k$

Scarti: $x_i - \overline{x}$

La somma totale degli scarti è sempre zero:

$\sum_{i=1}^{n}(x_i - \overline{x}) = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \overline{x} = n\overline{x} - n\overline{x} = 0$

Mediana campionaria: la media campionaria è influenzata dai valori outlier e per questo motivo è necessario trovare un indice statistico che non venga alterato da tali valori. In tal senso, ci viene in aiuto la mediana, definita come il valore intermedio quando i dati sono disposti in ordine crescente. La mediana è un indice di centralità robusto, in quanto non subisce perturbazioni dai valori outlier. Se il numero di valori è dispari allora la mediana campionaria è il valore intermedio della lista ordinata; se è pari è la media dei due valori intermedi

Moda campionaria: valore che si verifica con maggior frequenza nell'insieme di dati. Se non c'è un valore più frequente, tutti quelli con frequenza più alta sono detti *valori modali*

Varianza campionaria: le statistiche precedenti misurano il centro di un insieme di dati, la varianza, invece, misura la dispersione; per questo motivo si concentra sugli scarti

$s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$

Se traslassimo di una costante c tutti i valori x_i avremmo:

$y_i = x_i + c$

$y_i - \overline{y} = x_i + c - (\overline{x} + c) = x_i - \overline{x}$

Da ciò deduciamo che la varianza non è sensibile alla traslazione. Vediamo il caso con dilatazione:

$y_i = cx_i$

$s_y^2 = c^2 s_x^2$

Deviazione standard campionaria: radice quadrata positiva della varianza

$s = \sqrt{\frac{\sum_{i=1} n(x_i - \overline{x})}{n-1}}$

E nel caso di dilatazione:

$s_y = |c| s_x$

1.3 Note sugli insiemi di dati

Insieme di dati normale: un insieme di dati è definito normale se:

- Ha il punto di massimo nell'intervallo centrale
- Spostandosi verso destra o sinistra dal centro l'altezza diminuisce (forma a campana)
- Istogramma simmetrico rispetto all'intervallo centrale

Se l'istogramma è approssivamente simmetrico intorno alla mediana campionaria viene detto approssivamente normale. Se non rientra in questi due casi, l'istogramma è detto asimmetrico. L'insieme dei dati si dice "asimmetrico a destra" o "asimmetrico a sinistra" a seconda di quale sia la coda più lunga. L'istogramma potrebbe anche avere più di un massimo. Quando, in particolare, i massimi locali sono due, parleremo di istogramma bimodale.

Regola empirica: dato un insieme di dati approssivamente normale con media campionaria $\overline{x}$ e devazione standard campionaria s allora valgono:

1. Approssivamente il 68% delle osservazioni rientrano nell'intervallo $x \pm s$
2. Approssivamente il 95% delle osservazioni rientrano nell'intervallo $x \pm 2s$
3. Approssivamente il 99.7% delle osservazioni rientrano nell'intervallo $x \pm 3s$

Coefficiente di correlazione campionaria: statistica atta alla misurazione dell'associazione tra i valori di un insieme di dati a coppie. Considero la somma:

$\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$

Quando grandi valori di x tendono ad essere associati a grandi valori di y, e piccoli valori di x tendono ad essere associati a picoli valori di y, allora i segni positivi o negativi dei vari scarti tenderanno ad essere gli stessi. La sommatoria appena vista, divisa per n-1, prende il nome di **covarianza campionaria**

$Cov = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{n-1}$

La covarianza è un indice dimensionale, caratterizzata dall'unità di misura ottenuta dal prodotto fra l'unità di misura delle x e quella delle y. Se volessimo un indice adimensionale, divido la covarianza per il prodotto delle deviazioni standard di x e di y

$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{(n-1)s_x s_y}$

Tale indice prende il nome di **coefficiente di correlazione campionaria**. Se $r > 0$ i dati sono correlati positivamente, quando $r < 0$ i dati sono correlati negativamente

Proprietà:

1. r è sempre compreso fra -1 e +1
   - Posso inserirlo in una scala. Tanto più è vicino a 1 e tanto più la relazione tendenziale è forte, viceversa altrimenti
2. Se r è il coefficiente di correlazione campionaria di x_i e y_i, allora, scelti a, b, c e d a scelta, r sarà il coefficiente di correlazione campionaria anche per:

$a + bx_i$

$c + dy_i$

A condizione che b e d abbiano lo stesso segno ($bd >= 0$). Vediamo di dimostrare le due proprietà precedentemente enunciate:

Proprietà 1

$\forall i \quad y_i = a + bx_i \rightarrow \overline{y} = a + b\overline{x}$

Da cui:

$y_i - \overline{y} = a + bx_i - a - b\overline{x} = b(x_i - \overline{x})$

$r = \frac{1}{n-1}\frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{s_x s_y} = \frac{1}{n-1}\frac{b\sum(x_i - \overline{x})^2}{|b|s_x^2} = \frac{b}{|b|}\frac{s_x^2}{s_x^2} = \frac{b}{|b|}$

Quindi avremo che r sarà pari a 1 per $b > 0$ e a -1 per $b < 0$.

Proprietà 4

$(x_1, y_1), ..., (x_n, y_n)$

$\forall i \quad x'_i = a + bx_i \rightarrow \overline{x'} = a + b\overline{x} \rightarrow s_{x'} = |b|s_x \rightarrow x'_i - \overline{x'} = b(x_i - \overline{x})$

$\forall i \quad y'_i = c + dy_i \rightarrow \overline{y'} = c + d\overline{y} \rightarrow s_{y'} = |d|s_y \rightarrow y'_i - \overline{y'} = d(y_i - \overline{y})$

$r' = \frac{1}{n-1} \frac{\sum (x'_i - \overline{x'})(y'_i - \overline{y'})}{s'_x s'_y} = \frac{1}{n-1} \frac{bd \sum (x_i - \overline{x})(y_i - \overline{y})}{|b||d| s_x s_y} = \frac{bd}{|b||d|} r$

Tale quantità sarà pari a r se $bd > 0$ oppure a $-r$ se $bd < 0$

Formula alternativa per il calcolo di r:

$r = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{xy}}{\sqrt{(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2)(\sum_{i=1}^{n} y_i^2 - n\overline{y}^2)}}$

Quantili: supponendo di avere $x_1, x_2, ..., x_n$ e q appartenente all'intervallo chiuso $[0, 1]$, il quantile q-esimo è:

- $\geq$ di almeno nq osservazioni ordinate
- $\leq$ di almeno $n(1 - q)$ osservazioni ordinate

La mediana campionaria è un caso particolare di una statistica nota come 100p-esimo percentile campionario, dove p indica qualunque frazione compresa tra 0 e 1. Un 100p-esimo percentile campionario è un valore maggiore del 100p percento dei valori e minore del 100(1-p) di essi. Il percentile p-esimo è uguale al quantile p-esimo/100. Per calcolare il quantile q-esimo è necessario:

1. Disporre i dati in ordine crescente
2. Se nq non dovesse essere intero, è necessario determinare il più piccolo intero maggiore di nq. Il valore dei dati in questa posizione è il q-esimo quantile campionario
3. Se nq dovesse essere intero, allora la media dei valori nelle posizioni nq e $nq + 1$ è il q-esimo quantile campionario

Box Plot: rappresentazione grafica che riassume le principali caratteristiche di un campione di dati. Tale rappresentazione contiene due componenti principali:

- una *scatola*, intesa come un rettangolo che evidenzia il primo e il terzo quartile campionario dei dati, che corrispondono alle due basi, e la mediana, indicata tramite un segmento parallelo alle basi stesse;
- due *baffi*, che si estendono dagli estremi della scatola fino a raggiungere il minimo e il massimo valore osservato.

Il box è definito dal range interquantile, dato dalla differenza fra il III e il I quantile.

QQ Plot: rappresentazione grafica che considera due campioni al fine di valutare la validità dell'ipotesi che i campioni stessi seguano una medesima distribuzione. Questi diagrammi si basano sul fatto (che non dimostreremo) che i quantili campionari rappresentano l'approssimazione di quantili teorici che, considerati tutti insieme, individuano univocamente la distribuzione dei dati.

Pertanto, se due campioni hanno un'uguale distribuzione, allora estraendo da entrambi il quantile di un livello fissato si dovranno ottenere due numeri molto vicini (in quanto essi rappresentano approssimazioni diverse di uno stesso valore).

1.4 Eterogeneità/Omogeneità

Massima eteronegeità (minima omogeneità): gli elementi hanno tutti frequenza 1

Massima omogeneità (minima eterogeneità): contiene sempre la stessa forma ripetuta

Indice di Gini: dati m elementi $v_1, v_2, ..., v_m$ di frequenza $f_1, f_2, ..., f_m$, l'indice di Gini è definito come:

$I = 1 - \sum_{j=1}^{m} f_j^2$

$\forall j \to f_j \geq 0 \qquad \sum f_j = 1 \qquad \exists j | f_j > 0$

$f_j^2 > 0 \qquad \sum_j f_j^2 = f_j^2 + \sum_{j \neq \bar{j}} f_j^2 > 0$

$I = 1 - \sum_j f_j^2 < 1$

$0 \leq f_j \leq 1 \qquad \forall j \to f_j^2 \leq f_j$

$\sum_j f_j^2 \leq \sum_j f_j = 1 \to I = 1 - \sum_j f_j^2 \geq 1 - 1 = 0$

$0 \leq I < 1$

Nel caso di eterogeneità massima l'indice di Gini è: $I = \frac{m-1}{m}$

Nel caso di omogeneità massima l'indice di Gini è: $I = 0$

Dimostrazione

Massima eterogeneità

$\forall j \quad f_j = \frac{1}{m}$

$I = 1 - \sum_j^m f_j^2 = 1 - \sum_j^m \left(\frac{1}{m}\right)^2 =$

$1 - m\frac{1}{m^2} = 1 - \frac{1}{m} = \frac{m-1}{m}$

Massima omogeneità

$\exists \bar{j} \quad f_{\bar{j}} = 1$

$\forall j \neq \bar{j} \quad f_j = 0$

$I = 1 - \sum_j^m {f_j}^2 = 1 - 1 = 0$

Entropia

$H = \sum_j f_j log\frac{1}{f_j} \qquad H \geq 0$

Nel caso di eterogeneità massima l'entropia è: $H = log(m)$

Nel caso di omogeneità massima l'entropia è: $H = 0$

Concentrazione

La porzione di corso relativa alla concentrazione è da leggere sugli appunti

Trasformazioni lineari: fissate due costanti a e $b \in R$, il valore x verrà trasformato nel valore x' secondo la regola:

$x' = g(x) = ax + b$

Traslazione: $x \to x' = x \pm k$

Dilatazione: $x \to x' = \frac{x}{h}$

- $h > 1$ applica una contrazione, $h < 1$ applica una dilatazione
- Media, mediana, quantili, range di variazione, distanza interquantile e deviazione standard vengono scalati di $\frac{1}{h}$; la varianza viene scalata di $\frac{1}{h^2}$

Cambiamento di scala e di origine: se abbiamo dei valori nell'intervallo $[a, b]$ e vogliamo adattarli in modo che appartengano all'intervallo $[c, d]$, la trasformazione da applicare sarà:

$x \to x' = c + \frac{d-c}{b-a}(x - a)$

Standardizzazione: applico una scala il cui fattore è uguale alla deviazione standard dei valori, per poi traslare verso sinistra rispetto alla media. Il nuovo insieme dei valori avrà media 0 e varianza 1

$x' = \frac{x-\overline{x}}{s_x}$

Trasformazioni logaritmiche: quando i valori di una variabile osservata sono molto grandi oppure molto distanziati, conviene pensare a tali valori come potenza di una data base, ragionando in termini del relativo esponente. Ciò corrisponde ad applicare una trasformazione logaritmica del tipo:

$x \to x' = logx$

Le scelte per la base del logaritmo sono tendenzialmente 10 e 2. Nel caso in cui i valori siano molto distanti e caratterizzati da una distribuzione di frequenza unimodale fortemente asimmetrica, la trasformazione logaritmica permette di ottenere una distribuzione di frequenza più simmetrica.

1.5 Analisi della varianza

Dato un insieme di osservazione di un dato attributo, possiamo dividerlo in G gruppi diversi. Indichiamo con $n_1, n_2, ..., n_G$ la numerosità dei vari gruppi e con $n = n_1 + n_2 + ... + n_G$ il numero totale di osservazioni. Fissato $g \in \{1, ..., G\}$ e $i \in \{1, ..., n_G\}$, denotiamo con x_i^g il valore della i-esima osservazione nel gruppo g.

- Media campionaria: $\overline{x} = \frac{1}{n}\sum_{g=1}^{G}\sum_{i=1}^{n_g} x_i^g$
- Mediana campionaria: $\overline{x}^g = \frac{1}{n_g}\sum_{i=1}^{n_g}(x_i^g - \overline{x})^2 \qquad \forall g = 1, ..., n_G$
- Somma totale degli scarti: $SS_T = \sum_{g=1}^{G}\sum_{i=1}^{n_g}(x_i^g - \overline{x})^2$
- Somma degli scarti entro i gruppi (within groups): $SS_W = \sum_{g=1}^{G}\sum_{i=1}^{n_g}(x_i^g - \overline{x}^g)^2$
- Somma degli scarti tra i gruppi (between groups): $SS_B = \sum_{g=1}^{G} n_g(\overline{x}_i^g - \overline{x}^g)^2$
- Varianza campionaria su tutte le osservazioni: $s_T^2 = \frac{1}{n-1}SS_T$
- Varianza campionaria della media tra i gruppi: $s_B^2 = \frac{1}{G-1}SS_B$
- Varianca campionaria dei valori entro i gruppi: $s_W^2 = \frac{1}{n-G}SS_W$

Si può mostrare che:

$SS_T = SS_W + SS_B$

Che possiamo vedere come:

$\frac{SS_T}{n-1} = \frac{SS_W}{n-1} + \frac{SS_B}{n-1}$

$\frac{SS_T}{n-1} = \frac{n-G}{n-1}\frac{SS_W}{n-G} + \frac{G-1}{n-1}\frac{SS_B}{G-1}$

$s_T^2 = \frac{n-G}{n-1}s_W^2 + \frac{G-1}{n-1}s_B^2$

La dimostrazione di questa uguaglianza è lasciata al lettore

1.6 Analisi dei classificatori

Immaginiamo di avere a disposizione un classificatore *binario*, costruito cioè per discriminare tra due classi che convenzionalmente indicheremo come *positiva* e *negativa*. A partire da un insieme di oggetti di cui è noto a priori l'esito della classificazione (useremo la dicitura *oggetti positivi* e *oggetti negativi* per indicare gli oggetti che appartengono alle due classi), possiamo valutare la bontà di questo classificatore calcolando il numero di casi (o la corrispondente frazione) che vengono classificati in modo errato. Notiamo però che ci sono due possibili modi di sbagliare la classificazione:

- un esempio positivo viene classificato come negativo, dando luogo a un cosiddetto **falso negativo**
- un esempio negativo viene classificato come positivo, e in questo caso si parla di **falso positivo**

Tenendo conto del fatto che tipicamente è molto difficile riuscire a ottenere un buon classificatre in termini sia di falsi positivi, sia di falsi negativi, un modo efficace di valutare entrambi questi tipi di errore consiste nel disegnare la **matrice di confusione** (o *tabella di confusione*) che mostra il numero di falsi positivi e di falsi negativi unitamente al numero di casi correttamente classificati, a loro volta divisi in *veri positivi* e *veri negativi*.

A partire dalla matrice di confusione è possibile derivare due indici che valutano separatamente la capacità del classificatore a lavorare correttamente con gli oggetti positivi e con quelli negativi:

- la *sensibilità*, intesa come frazione degli oggetti positivi che vengono correttamente classificati

$Sensibilit = \frac{VP}{TP}$

- la *specificità*, analaogamente intesa come frazione degli oggetti negativi che vengono correttamente classificati

$Specificit = \frac{VN}{TN}$

Una volta calcolati i valori per questi due indici, è possibile valutare il classificatore in funzione della posizione assunta dal punto di coordinate $(1-Specificit, Sensibilit)$ sul piano cartesiano. In termini delle quantità sopra definite, le coordinate coincidono con $\left(1 - \frac{VN}{TN}, \frac{VP}{TP}\right)$, o equivalentemente con $\left(\frac{FP}{TN}, \frac{VP}{TP}\right)$. Vediamo nel seguito alcuni casi speciali.

Classificatori costanti

Consideriamo il classificatore CP che associa indiscriminatamente gli oggetti nella classe positiva.

Tutti i TP oggetti positivi verranno assegnati (correttamente) alla classe positiva, e tutti i TN oggetti negativi saranno assegnati (erroneamente) alla classe positiva. Ciò significa che il numero di veri positivi sarà pari a TP e il numero di veri negativi sarà zero: pertanto la sensibilità sarà uguale a 1 (com'è giusto che sia: tutti gli oggetti sono classificati come positivi e quindi il 100% degli oggetti positivi viene correttamente classificato) mentre la specificità sarà nulla (nessun oggetto negativo verrà classificato come tale). Il classificatore CP individuerà quindi il punto di coordinate $(1 - \text{specificità}, \text{sensibilità}) = (1, 1)$.

Il classificatore CN che associa tutti gli oggetti alla classe negativa si comporta in modo duale rispetto a CP.

In questo caso abbiamo che la sensibilità si annulla e la specificità vale 1, pertanto CN individuerà il punto $(0, 0)$ sul piano cartesiano.

Classificatori ideali

Un *classificatore ideale* è un classificatore in grado di non commettere alcun errore.

Un altro caso da considerare è quello in cui nessun punto verrà classificato correttamente, facendo sì che tutti gli oggetti considerati diano luogo ad errori di classificazione. Ci troviamo dunque di fronte a un classificatore totalmente errato, a partire dal quale è però sorprendentemente facile ottenere il classificatore ideale invertendo gli esiti della classificazione: ogni qual volta un oggetto verrebbe classificato come negativo lo si associa alla classe positiva e viceversa.

Classificatori casuali

Consideriamo il caso in cui sensibilità e specificità siano entrambe uguali a $\frac{1}{2}$, e dunque che metà degli oggetti positivi e metà di quelli negativi vengano classificati correttamente (il che significa che le due rimanenti metà vengono classificate male). Si può pertanto trarre la conclusione che dal punto di vista della sua bravura nell'assegnare gli oggetti alle due classi, questo classificatore è essenzialmente equivalente a un classificatore $\mathrm{CC}_{1/2}$ che assegna un generico oggetto a una classe scelta uniformemente a caso, per esempio lanciando una moneta.

Classificatore a soglia

Un *classificatore a soglia* effettua il procedimento di classificazione di un generico oggetto calcolando una quantità e verificando poi che quest'ultima sia superiore a una soglia prefissata. La quantità varierà ovviamente in funzione dell'oggetto considerato, mentre la soglia resterà uguale. Chiaramente, la costruzione di un tale tipo di classificatore richiede anche di fissare questo valore per la soglia. Gli indici di sensibilità e specificità possono essere utilizzati proprio per questo scopo: indicato con θ un generico valore per la soglia e individuato un intervallo $[\theta_min, \theta_max]$ in cui il valore può variare, si può considerare un'opportuna discretizzazione finita di tale intervallo (a meno che l'insieme dei valori considerabili non sia già discreto, finito e ragionevolmente piccolo) $D = \{\theta_0 = \theta_min, \ldots, \theta_n = \theta_max\}$. Per ogni $\theta \in D$ è poi possibile calcolare la sensibilità e la specificità del classificatore (va notato che è sufficiente costruire un'unica volta il classificatore a partire dai dati a disposizione, per poi variare via via la soglia) e disegnare sul piano cartesiano il punto corrispondente. Il risultato è una traiettoria che prende il nome di *curva ROC*.

L'andamento di una curva ROC ha sempre l'origine e il punto $(1, 1)$ come estremi. Infatti quando la soglia assume rispettivamente i suoi valori minimo e massimo il classificatore ha un output costante: nel primo caso tutti gli esempi saranno associati alla classe positiva, nel secondo a quella negativa. Quindi si ottengono i classificatori CP e CN che si collocano appunto in corrispondenza di $(0, 0)$ e $(1, 1)$ nel grafico. Inoltre la curva è sempre monotona non decrescente, e ciò è dovuto al fatto che all'aumentare della soglia il numero di oggetti classificati positivamente può solo decrescere.

Il valore di θ può quindi essere scelto in modo da trovare un giusto compromesso tra sensibilità e specificità. Il grafico della curva ROC viene inoltre utilizzato per valutare la bontà del classificatore indipendentemente da uno specifico valore della soglia. Ciò viene fatto misurando l'area compresa tra l'asse delle ascisse e la curva stessa.

Il valore di tale area viene indicato con la sigla AUC (che corrisponde ad "Area Under the ROC Curve"): più si avvicina a 1, più il classificatore ha un comportamento che approssima quello del caso ideale CI.

1.7 Calcolo combinatorio

Permutazioni: dato un insieme di n oggetti $A = \{a_1, a_2, ..., a_n\}$, una permutazione di tali oggetti è una qualsiasi sequenza ordinata in cui compaiono tutti gli oggetti.

Se gli n oggetti di A sono tutti distinguibili si parla di permutazioni semplici, calcolabili come:

$P_n = n \times (n-1) \times (n-2)...2 \times 1 = n!$

Permutazioni di oggetti distinguibili a gruppi: quando gli oggetti di A non sono tutti distinguibili ma sono distinguibili a gruppi di numerosità $n_1, n_2, ..., n_k$ (la cui somma ovviamente vale n) allora una sequenza ordinata di tali oggetti che sia distinguibile dalle altre è detta permutazione di oggetti distinguibili a gruppi. Formalizziamo:

$P_{n;n_1,...,n_k} = \frac{n!}{n_1!n_2!...n_k!} = \binom{n}{n_1!n_2!...n_k!}$

Tale coefficiente è chiamato *coefficiente multinomiale*.

Disposizioni e combinazioni: consideriamo n oggetti distinti di un insieme $A = \{a_1, a_2, ..., a_n\}$ e selezioniamone k.

- Se vogliamo distinguere le configurazioni contenenti gli stessi oggetti ma estratti in ordine differente, allora parliamo di disposizioni di n oggetti su k posti, dove sono importanti sia l'oggetto selezionato che la sua posizione
- Se siamo interessati a quali oggetti sono stati estratti e non alla loro posizione nella sequenza, parliamo di combinazioni di n oggetti presi k alla volta

Parliamo di disposizioni o combinazioni senza ripetizione se gli oggetti di A possono essere usati una sola volta. Se il singolo oggetto può essere selezionato anche più di una volta, allora parliamo di disposizioni o combinazioni con ripetizione.

Dispozioni senza ripetizione

$d_{n,k} = n(n-1)(n-2)...(n-k+1) = n(n-1)(n-2)...(n-k+1)\left(\frac{(n-k)(n-k-1)...1}{(n-k)(n-k-1)...1}\right) = \frac{n!}{(n-k)!}$

Combinazioni senza ripetizione

$c_{n,k} = \left(\frac{d_{n,k}}{k}\right)! = \frac{n!}{(n-k)!k!} = \binom{n}{k}$

Disposizioni con ripetizione

$D_{n,k} = n \times n... \times n = n^k$

Combinazioni con ripetizione

$C_{n,k} = \binom{n+k-1}{k}$

1.8 Teoria degli Insiemi

Identifichiamo con Ω l'*insieme Universo*, ossia l'insieme che contiene tutti gli elementi possibili. Diremo che $w \in \Omega$ quando l'elemento w appartiene all'insieme Ω. Definiamo $E \subseteq \Omega$ quando E è un sottoinsieme dell'insieme Ω.

Identifichiamo infine con $\emptyset$ l'insieme vuoto e con $\{\}$ l'insieme privo di elementi.

Unione: $E \cup F \qquad x \in E \cup F \leftrightarrow x \in E \vee x \in F$

Intersezione: $E \cap F \qquad x \in E \cap F \leftrightarrow x \in E \wedge x \in F$

Disgiunzione: $E \cap F = \emptyset$

Complemento di E: E^C oppure $\overline{E}$

Notazioni particolari:

- $\bigcap_{i=1}^{n} A_i = A_1 \cap A_2 ... \cap A_n$
- $\bigcup_{i=1}^{n} A_i = A_1 \cup A_2 ... \cup A_n$

Proprietà

- *Commutative*

$E \cap F = F \cap E \qquad E \cup F = F \cup E$

- *Associative*

$(E \cap F) \cap G = \quad E \cap (F \cap G)$

- *Distributive*

$(E \cup F) \cap G = (E \cap G) \cup (F \cap G)$

$(E \cap F) \cup G = (E \cup G) \cap (F \cup G)$

- *De Morgan*

$\overline{E \cup F} = \overline{E} \cap \overline{F}$

$\overline{E \cap F} = \overline{E} \cup \overline{F}$

1.9 Elementi di probabilità

Interpretazione frequentista: la probabilità di un esito è considerata una proprietà dell'esito stesso. In tal caso, essa viene calcolata come rapporto fra il numero di casi favorevoli e il numero di casi possibili

Interpretazione soggettivistica: la probabilità di un esito non è una proprietà oggettiva, bensì la precisazione del livello di fiducia che lo studioso ripone nel verificarsi di un evento.

L'insieme universo Ω viene spesso chiamato anche "*spazio campionario*" oppure "*spazio degli eventi*": in entrambi i casi essi identificano l'insieme di tutti gli esiti di un esperimento casuale. Definiamo "*evento elementare*" un elemento w appartenente all'universo Ω. Definiamo banalmente "*evento*" un generico sottoinsieme dello spazio degli eventi: $E \subset \Omega$. L'evento può essere interpretato come un sottoinsieme contenente eventi elementari. Gli insiemi composti da un solo elemento sono detti *insiemi singoletto*. Tra tutti gli insiemi ve ne sono due di menzione particolare:

- {}: l'insieme di partenza, si verifica sempre
- $\emptyset$: l'insieme vuoto, non si verifica mai

Proprietà:

- Disgiunzione logica: evento che si verfica quando si verifica almeno uno degli eventi che ho unito

$E_1 \cup E_2 ... \cup E_n$

- Congiunzione logica: evento che si verifica quando sia E che F si sono verificati

  $E_1 \cap E_2$

- Evento certo: l'universo Ω è l'evento che contiene tutti gli esiti possibili. Ω si verifica sempre

  $E \subseteq \Omega$

- Evento impossibile: qualunque esito dell'esperimento io analizzi, non sarà mai nell'insieme vuoto, quindi in questo caso si parla di evento impossibile

  $\emptyset$

- $\overline{E} = \Omega \setminus E \qquad \overline{\Omega} = \emptyset \rightarrow \overline{\emptyset} = \Omega$
- $E \subseteq F$ E implica F
- $E \subseteq F$ e $F \subseteq E \rightarrow E = F$

Algebra degli eventi: definiamo Algebra degli eventi l'insieme:

$A = \{E_i \subseteq \Omega\}$

Per essere tale, è necessario che vengano soddisfatte le seguenti proprietà:

1. $\Omega \in A$
2. $\forall E$ se $E \in A$ allora $\overline{E} \in A$
3. $\forall E, F$ se $E \in A \land F \in A$ allora $E \cup F \in A$
4. L'algebra è chiusa rispetto all'unione infinita. Per questo motivo A è più di un algebra e viene detta σ-algebra

Funzione di probabilità: $P : A \rightarrow [0,1]$

Dominio: generica algebra A

Codominio: tutti i valori dell'intervallo chiuso [0,1]

Assiomi di Kolmogorov

1. $P(\Omega) = 1$ posso calcolare tale probabilità perché $\Omega \in A$ per ipotesi
2. $\forall E \in A \quad 0 \leq P(E) \leq 1$ probabilità intesa come frequenza relativa con cui a lungo termine si verifica un determinato evento
3. $\forall E_1, E_2 \in A$ se $E1 \cap E2 = \emptyset$ diremo che $E1$ e $E2$ sono eventi disgiunti o *mutuamente esclusivi* e in tal caso vale che:

$P(E_1 \cup E_2) = P(E_1) + P(E_2)$

4. $\forall E_1, ..., E_n \in A \qquad \forall i \neq j \quad E_i \cap E_j = \emptyset$

$P(\bigcup_{i=1}^{n} E_i) = \sum_{i=1}^{n} P(E_i)$

Teoremi elementari di probabilità

1. $\forall E \in A \quad P(\overline{E}) = 1 - P(E)$

   Dimostrazione

   $1 = P(\Omega) = P(E \cup \overline{E}) = P(E) + P(\overline{E}) \quad \rightarrow \quad P(\overline{E}) = 1 - P(E)$

2. $\forall E, F \in A \quad P(E \cup F) = P(E) + P(F) - P(E \cap F)$

Spazi equiprobabili

(Ω, A, P) = Spazio di probabilità

Supponiamo Ω composto da N eventi: $\Omega = \{1, 2, ..., N\}$; se tutti gli eventi hanno la stessa probabilità, parliamo di spazi equiprobabili.

$\forall w \in \Omega \qquad P(\{w\}) = p$

$\Omega = \{1\} \cup \{2\} ... \cup \{N\} \qquad 1 = P(\Omega) = P(\{1\}) + ... + P(\{N\}) = Np \to p = \frac{1}{N}$

N.B. E' necessario che Ω sia finito, poiché altrimenti N tenderebbe a infinito e, di conseguenza, p tenderebbe a 0. Se consideriamo ora un sottoinsieme di Ω, abbiamo:

$\forall E \subseteq \Omega \qquad E = \{e_1, ..., e_k\} \qquad k \leq N$

$E = \{e_1\} \cup \{e_2\} ... \cup \{e_k\}$

$P(E) = P(\{e_1\}) + ... + P(\{e_k\} = kp = \frac{k}{N} = \frac{|E|}{N}$

Probabilità condizionata

$P(E|F) = \frac{P(E \cap F)}{P(F)} \qquad$ vale se $P(F) \neq 0$

Teorema delle probabilità totali

Supponiamo di avere due eventi E, $F \in A$ con:

$(E \cap F) \cup (E \cap \overline{F}) = E$

$P(E) = P(E|F)P(F) + P(E|\overline{F})P(\overline{F}) = P(E|F)P(F) + P(E|\overline{F})(1 - P(F))$

Partizionamento

Supponiamo di avere l'insieme Ω partizionato in n sezioni di stessa grandezza: $F_1, F_2, ..., F_n$. L'unione di queste partizioni costruisce lo spazio degli eventi. Inoltre: $\forall i \neq j \to F_i \cap F_j = \emptyset$

Supponiamo di analizzare un evento E e di osservare le intersezioni fra tale evento e tutte le partizioni dello spazio degli eventi: $E \cap F_1, E \cap F_2, ..., E \cap F_n$

$\bigcup_{i=1}^{n} (E \cap F_i) = E \cap (\bigcup_{i=1}^{n} F_i) = E \cap \Omega = E$

$\forall i \neq j \to (E \cap F_i) \cap (E \cap F_j) = (E \cap E) \cap (F_i \cap F_j) = E \cap \emptyset = \emptyset$

Dall'unione di quanto appena enunciato, posso dire che:

$E = E \cap (\bigcup_{i=1}^{n} F_i) \to P(E) = \sum_{i=1}^{n} P(E \cap F_i) = \sum_{i=1}^{n} P(E|F_i)P(F_i)$

Teorema di Bayes

Immaginiamo di avere n eventi che soddisfino le ipotesi del problema delle probabilità totali:

$F_1, F_2, ..., F_n$

$P(F_j|E)P(E) = P(F_j \cap E) = P(E \cap F_j) = P(E|F_j)P(F_j)$

$P(F_j|E) = \frac{P(E|F_j)P(F_j)}{P(E)} = \frac{P(E|F_j)P(F_j)}{\sum_{i=1}^{n} P(E|F_i)P(F_i)}$

Eventi indipendenti

Nel caso in cui $P(E|F)$ e $P(E)$ siano uguali, diciamo che E ed F sono indipendenti. Quindi, se:

$P(E|F) = \frac{P(E \cap F)}{P(F)}$

Allora, possiamo formalizzare l'indipendenza di eventi dicendo che:

$P(E \cap F) = P(E)P(F)$

Una proprietà da notificare è che se E ed F sono indipendenti, lo sono anche E e $\overline{F}$

Supponiamo ora di avre più eventi: E, F e G. Tali eventi si dicono indipendenti se:

1. $P(E \cap F) = P(E)P(F)$
2. $P(E \cap G) = P(E)P(G)$
3. $P(F \cap G) = P(F)P(G)$
4. $P(E \cap F \cap G) = P(E)P(F)P(G)$

Generalizzando, presi n eventi diversi, essi si dicono indipendenti se e solo se per ogni sottogruppo $E_{\alpha 1}, ..., E_{\alpha r}$, vale l'equazione:

$P(\bigcap_{i=1}^{r} E_{\alpha i}) = \Pi_{i=1}^{r} P(E_{\alpha i})$

Se tre o più eventi sono indipendenti, allora ciascuno di essi è indipendente da qualunque evento si possa costruire con gli altri due.

$P(E \cap (F \cup G)) = P(E \cap F) \cup P(E \cap G) = P(E \cap F) + P(E \cap G) - P(E \cap F \cap G) = P(E)P(F) + P(E)P(G) - P(E)P(F \cap G) = P(E)(P(F) + P(G) - P(F \cap G)) = P(E)P(F \cup G)$

Componenti in serie e in parallelo

Immaginiamo un sistema di n scatole poste in serie. Tali scatole conducono elettricità, quindi è necessario che siano tutte funzionanti affinché il sistema funzioni correttamente:

$P(Scatola_i = 1) = p_i$

Ogni scatola è indipendente dalle altre, quindi:

$P(sistemaFunziona) = P(\bigcap_{i=1}^{n} Scatola_i = 1) = \Pi_{i=1}^{n} P(Scatola_i = 1) = \Pi_{i=1}^{n} p_i$

Ora immaginiamo di prendere quelle scatole e di porle in parallelo. Così facendo, il sistema è più robusto ai danni, poiché anche se una componente dovesse rompersi, ne avrei ancora $n-1$ su cui fare affidamento. Fintanto che c'è almeno un componente funzionante, il sistema funzionerà, quindi:

$P(sistemaFunziona) = 1 - P(sistemaRotto) = 1 - P(\bigcap_{i=1}^{n} Scatola_i = 0) = 1 - \Pi_{i=1}^{n} P(Scatola_i = 0) = 1 - \Pi_{i=1}^{n}(1 - p_i)$

Naive-Bayes

Supponiamo di avere una situazione ove vengono associati dei particolari valori a dei determinati attributi (e.g. occhi = attributo; nero = valore). Indichiamo:

- Attributi: $X_1, X_2, ..., X_n$
- Valori: $X_1 = x_1, X_2 = x_2, ..., X_n = x_n$

Ora, indichiamo con $Y = y_k$ il generico k-esimo elemento di un insieme di cui ci interessa studiarne gli attributi. Abbiamo:

$P(Y = y_k | X_1 = x_1, ..., X_n = x_n) \rightarrow$ Applichiamo Bayes $\rightarrow \frac{P(X_1 = x_1, ..., X_n = x_n | Y = y_k) P(Y = y_k)}{P(X_1 = x_1, ..., X_n = x_n)}$

Da cui:

$\frac{P(X_1=x_1|Y=y_k)...P(X_n=x_n|Y=y_k)P(Y=y_k)}{P(X_1=x_1,...,X_n=x_n)}$

Questo rapporto è proporzionale a:

$\Pi_{i=1}^{n} P(X_i = x_i|Y = y_k)P(Y = y_k)$

Chiamiamo *classificatore Naive-Bayes* il seguente valore:

$k^* = argmax_k \Pi_{i=1}^{n} P(X_i = x_i|Y = y_k)P(Y = y_k)$

Variabili aleatorie

Definiamo *variabile aleatoria* una qualsiasi variabile:

$X : \Omega \to R$

Scrivendo $X = \alpha$ indichiamo l'evento $X = \alpha$, che formalizzando sarebbe:

$\{X = \alpha\} = \{w \in \Omega : X(w) = \alpha\}$

Per semplicità, anziché scrivere $P(\{X = \alpha\})$ scriveremo $P(X = \alpha)$.

Funzione indicatrice

Definiamo *funzione indicatrice* la funzione:

$$I = \begin{cases} 1 \\ 0 \end{cases}$$

che assume valore 1 quando si verifica un determinato evento, 0 altrimenti.

Funzione di ripartizione

Prendiamo una variabile aleatoria X e definiamo:

$F_X : R \to [0, 1]$

$F_X(x) = P(X \leq x)$

$\{X \leq b\} = \{X \leq a\} \cup \{a < X \leq b\}$ per $a < b$

Da cui deriviamo:

$P(X \leq b) = P(X \leq a) + P(a < x \leq b)$

Applicando la sostituzione otteniamo:

$F_X(b) = F_X(a) + P(a < X \leq b)$

$P(a < X \leq b) = F_X(b) - F_X(a)$

Variabili aleatorie discrete

Le variabili aleatorie discrete assumono un insieme numerabile di specificazioni. Introduciamo la *funziona di massa di probabilità* definendola come:

$p_X : R \to [0, 1]$

$\forall x \in R \qquad p_X(x) = P(X = x)$

Proprietà:

1. $p_X \neq 0$ in un insieme numerabile
2. $p_X \geq 0$
3. $\sum_i p_X(x_i) = 1$

Possiamo definire la *funzione di ripartizione* in funzione di quella di massa di probabilità e viceversa:

$F_X(x) = P(X \leq x) = \sum_{a \leq X} P(X = a) = \sum_{a \leq X} p_X(a)$

Proprietà

1. $F_X(x) \geq 0 \qquad \forall x \in R$
2. $lim_{x \to +\infty} F_X(x) = 1 \qquad\qquad lim_{x \to +\infty} P(X \leq x)$
3. F_X è continua da destra

Valore atteso

Dato la variabile aleatoria discreta X su $\{x_1, x_2, ..., x_n\}$ e p_X come funzione di massa di probabilità, definisco valore atteso la quantità:

$E(X) = \sum_i x_i p(x_i) = \sum_i x_i P(X = x_i)$

Come possiamo vedere dalla formula soprastante, il valore atteso è una quantità dimensionale, che assume la stessa dimensione di x.

Consideriamo ora:

$g(X) = ax + b$ con $a, b \in R$

$X \to g(X) = aX + b$

$E(g(X)) = \sum_i (ax_i + b)P(X = x_i) = \sum_i ax_i P(X = x_i) + \sum_i bP(X = x_i) = a\sum_i x_i P(X = x_i) + b\sum_i P(X = x_i) = aE(X) + b$

Quindi:

$E(aX + b) = aE(X) + b$

1. Se $a = 0 \to E(b) = b$
2. Se $b = 0 \to E(aX) = aE(X)$

Varianza

Consideriamo la variabile aleatoria X e il suo valore atteso $E(X) = \mu$. Se considero:

$|X - \mu|$

difatti sto operando con una nuova variabile aleatoria, di cui posso calcolare il valore atteso:

$E(|X - \mu|)$

Ora, per togliere il valore assoluto ed operare esclusivamente con valori positivi, elevo al quadrato, ottenendo:

$E((X - \mu)^2)$

Questo nuovo valore prende il nome di *varianza* e viene indicata con σ_X^2

$Var(X) = E(X^2 - 2\mu X + \mu^2) = E(X^2) - 2\mu E(X) + \mu^2 = E(X^2) - 2\mu^2 + \mu^2 = E(X^2) - \mu^2 = E(X^2) - E(X)^2$

Proprietà:

1. $Var(aX + b) = E((aX + b - E(aX + b))^2) = E((aX + b - a\mu - b)^2) = E(a^2(X - \mu)^2) = a^2 E((X - \mu)^2) = a^2 Var(X)$
2. $Var(aX + b) \rightarrow$ se $a = 0 \rightarrow Var(b) = 0$

La varianza ha la stessa unità di misura della variabile aleatoria elevata al quadrato. Se vogliamo una statistica che mantenga l'unità di misura della variabile aleatoria utilizziamo la *deviazione standard*, calcolata come:

$\sigma_x = \sqrt{Var(X)}$

Variabile aleatoria multivariata

Ragioniamo in termini di coppie di variabili aleatorie: X, Y.

- *Funzione di ripartizione congiunta*: $F_{X,Y}(x, y) = P(X \leq x, Y \leq y)$
  - Nel caso in cui uno dei due valori tendesse a infinito, avremo: $lim_{y \rightarrow +\infty} F_{X,Y}(x, y) = lim_{y \rightarrow +\infty} P(X \leq x, Y \leq y) = P(X \leq x) = F_X(x)$
    - Tale valore è detto *funzione di ripartizione marginale*
- *Funzione di massa di probabilità congiunta*: $p_{X,Y}(x_i, y_j) = P(X = x_i, Y = y_j)$
  - Se fisso una delle due variabili e muovo l'altra: $\sum_j P(X = x_i, Y = y_j) = P(X = x_i) \rightarrow \sum_j p_{X,Y}(x_i, y_j) = p_X(x_i)$
    - Tale valore è detto *funzione di massa di probabilità marginale*

Diciamo che due variabili aleatorie X, Y sono indipendenti se e solo se:

$\forall A, B \subseteq R \rightarrow P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$

Questa formula vale anche nei seguenti casi:

1. $P(X \leq x, Y \leq y) = P(X \leq x)P(Y \leq y)$
2. $F_{X,Y}(x, y) = F_X(x)F_Y(y)$
3. $p_{X,Y}(x, y) = p_X(x)p_Y(y)$

La dimostrazione è lasciata al lettore.

Valore atteso di funzione di più variabili

Supponiamo di avere due variabili aleatorie X, Y e una funzione $g : R^2 \rightarrow R$

Avremo:

$E(g(X, Y)) = \sum_{x,y} g(x, y)P(X = x, Y = y) = \sum_{x,y} g(x, y)p_{X,Y}(x, y)$

Se prendessi $g(X, Y) = X + Y$ avrei $E(X + Y) = E(X) + E(Y)$

La dimostrazione è lasciata al lettore.

Per un generico n numero di variabili aleatorie avrei:

$E(\sum_{i=1}^n X_i) = \sum_{i=1}^n E(X_i)$

Covarianza

Date due variabili aleatorie X, Y e ponendo $\mu_x = E(X)$ e $\mu_y = E(Y)$, definisco *covarianza* la quantità:

$Cov(X, Y) = E((X - \mu_x)(Y - \mu_y))$

Proprietà:

1. $Cov(X, Y) = Cov(Y, X)$
2. $Cov(X, Y) = E(XY - \mu_x Y - \mu_y X + \mu_x \mu_y) = E(XY) - E(X)E(Y)$
3. $Cov(aX, Y) = E(aXY) - E(aX)E(Y) = aCov(X, Y)$
4. $Cov(X + Y, Z) = Cov(X, Z) + Cov(Y, Z) \rightarrow Cov(\sum_i X_i, Z) = \sum_i Cov(X_i, Z)$
5. $Cov(\sum_i \alpha_i X_i, Z) = \sum_i \alpha_i Cov(X_i, Z)$
6. $Cov(\sum_i \alpha_i X_i, \sum_j \beta_j Y_j) = \sum_i \sum_j \alpha_i \beta_j Cov(X_i, Y_j)$
7. $Cov(X, X) = Var(X)$
8. $Cov(X + b, Y) = Cov(X, Y)$
9. $Var(X + Y) = Var(X) + Var(Y) + 2Cov(X, Y)$
10. $Var(X + X) = Var(X) + Var(X) + 2Cov(X, X) = 4Var(X)$
11. $Var(\sum_i X_i) = \sum_{i,j} Cov(X_i, Y_j)$

Nel caso in cui le variabili aleatorie X, Y fossero indipendenti, avremmo che il valore atteso del loro prodotto è il prodotto dei valori attesi: $E(XY) = E(X)E(Y)$. Ma allora, la covarianza di due variabili aleatorie indipendenti è 0.

$Cov(X, Y) = E(XY) - E(X)E(Y) = E(X)E(Y) - E(X)E(Y) = 0$

E la varianza, di conseguenza, diventa:

$Var(X + Y) = Var(X) + Var(Y) + 2Cov(X, Y) = Var(X) + Var(Y)$

Generalizzando:

$Var(\sum_i X_i) = \sum_i Cov(X_i)$

Correlazione

Consideriamo $A, B \subseteq \Omega$ con $X = I_A, E(X) = P(X = 1)$ e $Y = I_B, E(Y) = P(Y = 1)$. Allora:

$XY = 1$ se $X = 1 \wedge Y = 1$

$XY = 0$ altrimenti

Il valore atteso delle due variabili sarà:

$E(XY) = P(X = 1, Y = 1)$

Mentre la covarianza:

$Cov(X, Y) = E(XY) - E(X)E(Y) = P(X = 1, Y = 1) - P(X = 1)P(Y = 1)$

$Cov(X, Y) > 0 \leftrightarrow P(X = 1, Y = 1) > P(X = 1)P(Y = 1)$

$\frac{P(X=1,Y=1)}{P(Y=1)} > P(X = 1) \leftrightarrow P(X = 1|Y = 1) > P(X = 1)$

Da ciò arriviamo all'*indice di correlazione lineare*:

$\rho_{X,Y} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$

Proprietà:

1. $\rho_{X,Y} \in [-1, 1]$
2. Se X, Y sono variabili aleatorie indipendenti, allora $Cov = 0 \rightarrow \rho_{X,Y} = 0$
3. Se ρ è -1 oppure 1 significa che tra X e Y c'è una dipendenza di tipo deterministico e, quindi, lineare
4. $\rho_{2X,2Y} = \frac{Cov(2X,2Y)}{\sigma_{2X}\sigma_{2Y}} = \rho_{X,Y}$

Variabili aleatorie continue

Una variabile aleatoria X è detta continua se:

1. $f_x : R \rightarrow R^+$

$\forall B \subseteq R$

2. $P(X \in B) = \int_B f_X(x)dx$.

Definiamo infine *densità di probabilità* la f_X.

Quindi abbiamo:

1. $P(a \leq X \leq b) = P(X \in [a, b]) = \int_a^b f_X(x)dx$
2. $P(X \in R) = \int_{-\infty}^{+\infty} f_X(x)dx$
3. $P(X = \alpha) = \int_\alpha^\alpha f_X(x)dx = 0$
4. $F_X(a) = P(X \leq a) = \int_{-\infty}^a f_X(x)dx$
5. $P(a - \frac{\epsilon}{2} \leq X \leq a + \frac{\epsilon}{2}) = \int_{a-\frac{\epsilon}{2}}^{a+\frac{\epsilon}{2}} f_X(x)dx \approx \epsilon f_X(a)$

Valore atteso

$E(X) = \int_{-\infty}^{+\infty} x f_X(x)dx$

Varianza

$Var(X) = E((X - \mu_X)^2)$

La funzione di ripartizione di due variabili aleatorie X, Y è uguale al caso discreto, da cui, se $X \geq 0$:

$E(X) = \int_0^{+\infty} 1 - F_X(x)dx$

Disuguaglianza di Markov

Supponendo $X \geq 0$, allora $\forall a > 0$ abbiamo:

$P(X \geq a) \leq \frac{E(X)}{a}$

Questa disuguaglianza vale sia nel caso discreto che continuo. La dimostrazione è lasciata al lettore. Cosa accade nel caso in cui abbiamo $P(X < a)$?

$P(X < a) = 1 - P(X \geq a) \geq 1 - \frac{E(X)}{a}$

Disuguaglianza di Chebyshev

Presa una variabile aleatoria X di valore atteso $E(X) = \mu$ e $Var(X) = \sigma^2$, allora $\forall r > 0$:

$P(|X - \mu| \geq r) \leq \frac{\sigma^2}{r^2}$

Vale sia per variabili aleatorie discrete che continue. Cosa accade nel caso in cui $P(|X - \mu| < r)$?

$P(|X - \mu| < r) = 1 - P(|X - \mu| \geq r) \geq 1 - \frac{\sigma^2}{r^2}$

L'importanza di queste due disequazioni risiede nel fatto che ci consentono di avere stime della probabilità a partire dal valore atteso o dall'unione del valore atteso con la varianza.

1.10 Modellizzazione discreta

Abbiamo bisogno di astrarre esperimenti casuali simili che si presentano in spoglie diverse. Partiamo analizzando gli esperimenti bernoulliani.

Modello di Bernoulli

Un esperimento bernoulliano è caratterizzato da una variabile aleatoria di bernoulli, parametrizzata rispetto alla probabilità di successo. Tale variabile aleatoria assume quindi due valori:

$X = 1$ se l'esito è positivo

$X = 0$ se l'esito è negativo

Definiamo distribuzione bernoulliana, la distribuzione identificata da:

$X \sim B(p)$

da leggersi "la variabile aleatoria X è distribuita come B(p)".

Funzione di massa di probabilità

$p_X(0) = P(X = 0) = 1 - p$

$p_X(1) = P(X = 1) = p$

$p_X(x) = 0 \qquad \forall x \neq 0 \wedge x \neq 1$

Generalizzando:

$p_X(x) = p^x(1-p)^{1-x} I_{\{0,1\}}(x)$

Funzione di ripartizione

$F_X(x) = P(X \leq x)$

Considerando che X può assumere solo valori pari a 0 o 1, abbiamo:

$F_X(0) = P(X \leq 0) = P(X = 0) = 1 - p \qquad F_x(1) = P(X \leq 1) = 1$

Generalizzando:

$F_X(x) = 0 I_{R^-}(x) + (1-p) I_{[0,1)}(x) + 1 I_{[1,+\infty)}(x) = (1-p) I_{[0,1)}(x) + 1 I_{[1,+\infty)}(x)$

Valore atteso

$E(X) = \sum_x x P(X = x) = 0 P(X = 0) + 1 P(X = 1) = P(X = 1) = p$

Varianza

$Var(X) = E((X-\mu)^2) = E((X-p)^2) = \sum_x (x-p)^2 P(X = x) = (0-p)^2 P(X = 0) + (1-p)^2 P(X = 1) = p^2(1-p) + (1-p)^2 p = p(1-p)(p+1-p) =$
$= p(1-p)$

La varianzasi massimizza per $p = \frac{1}{2}$

Vediamo alcuni casi estremi:

$p = 1 \to E(X) = 1 \quad Var(X) = 0$

$p = 0 \to E(X) = 0 \quad Var(X) = 0$

Per valori intermedi abbiamo che il valore medio e p crescono proporzionalmente.

Modello binomiale

Supponiamo di eseguire un esperimento bernoulliano per n volte. Otterremo una nuova variabile aleatoria chiamata *variabile aleatoria binomiale*. Essa ha due parametri:

- La variabile aleatoria su cui si basa
- Il numero di ripetizioni dell'esperimento bernoulliano

Formalizziamo:

$X \sim B(n, p)$

Con $n \in N$ e $p \in [0, 1]$. Il dominio di X sarà $D_X = \{0, ..., n\}$

Funzione di massa di probabilità

$p_X(i) = P(X = i) = \binom{n}{i} p^i (1-p)^{n-1} I_{\{0,...,n\}}(i)$

Funzione di ripartizione

$F_X(x) = P(X \leq x) = \sum_{i \leq x} P(X = i) = \sum_{i=0}^{\lfloor x \rfloor} P(X = i) = \sum_{i=0}^{\lfloor x \rfloor} \binom{n}{i} p^i (1-p)^{n-i} = \left(\sum_{i=0}^{\lfloor x \rfloor} \binom{n}{i} p^i (1-p)^{n-i} \right) I_{[0,n]}(x) + I_{[n,+\infty)}(x)$

Valore atteso

Possiamo vedere la nostra distribuzione $X \sim B(n, p)$ come $X_1, ..., X_n \sim B(p)$, da cui:

$X = \sum_{i=1}^{n} X_i$

Quindi:

$E(X) = E(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} E(X_i) = \sum_{i=1}^{n} p = np$

Quindi il valore atteso cresce linearmente al numero di prove effettuate.

Varianza

$Var(X) = Var(\sum_{i=1}^{n} X_i) \to$ assumendo X_i indipendenti $\to \sum_{i=1}^{n} Var(X_i) = \sum_{i=1}^{n} p(1-p) = np(1-p)$

Più aumenta n e più aumenta l'intervallo possibile di valori su cui può variare.

Cosa accade se cerchiamo di sommare due variabili aleatorie X_1 e X_2 che seguono il modello binomiale?

$X_1 \sim B(n, p)$

$X_2 \sim B(m, p)$

Abbiamo che:

$X_1 = \sum_{i=1}^{n} X_{1,i}$

$X_2 = \sum_{i=1}^{m} X_{2,i}$

Quindi:

$X_1 + X_2 = \sum_{i=1}^{n+m} X_i \sim B(n+m, p)$

Modello uniforme discreto

Supponiamo ora di non essere più nel caso bernoulliano. Abbiamo uno spazio degli eventi equiprobabili: cerchiamo di modellarlo con le variabili aleatorie.

$n \in N \setminus \{0\}$ è il numero di esiti equiprobabili

$X \sim U(n)$

$P(X = 1) = P(X = 2) = ... = P(X = n) = \frac{1}{n}$

Funzione di massa di probabilità

$P(X = x) = \frac{1}{n} I_{\{1,...,n\}}(x) = p_X(x)$

Funzione di ripartizione

$F_X(x) = P(X \leq x) = \sum_{i \leq x} p_X(i) = \sum_1^{\lfloor x \rfloor} \frac{1}{n} = \frac{\lfloor x \rfloor}{n} I_{[1,n]}(x) + I_{(n,+\infty)}(x)$

Valore atteso

$E(X) = \sum_{x=1}^n xP(X = x) = \sum_{x=1}^n x\frac{1}{n} = \frac{1}{n}\sum_{x=1}^n x = \frac{1}{n}\frac{n(n+1)}{2} = \frac{n+1}{2}$

Varianza

$E(X^2) = \sum_{x=1}^n x^2\frac{1}{n} = \frac{1}{n}\sum_{x=1}^n x^2 = \frac{1}{n}\frac{(n+1)(2n+1)}{6} = \frac{(n+1)(2n+1)}{6}$

$Var(X) = E(X^2) + E(X)^2 = \frac{n^2-1}{12}$

La dispersione intorno al valore centrale aumenta all'aumentare del numero di eventi equiprobabili.

Modello geometrico

La *distribuzione geometrica* descrive il numero di insuccessi necessari affinché si verifichi il primo successo in una successione di esperimenti bernoulliani indipendenti e identicamente distribuiti. Questa distribuzione, che ha quindi come supporto l'insieme dei numeri naturali (zero incluso), è quindi completamente descritta specificando il parametro p del corrispondente esperimento bernoulliano. Più precisamente, $p \in (0, 1]$: il caso $p = 0$ va infatti escluso a priori, altrimenti sarebbe impossibile avere come esito un successo (e dunque sarebbe impossibile contare il numero di insuccessi prima che si verifichi un successo). Il caso $p = 1$ può essere incluso nell'insieme dei valori validi per il parametro, sebbene questa scelta identifichi un esperimento bernoulliano che ha sempre successo: in tal caso, la variabile aleatoria che conteggia il numero di insuccessi prima del primo successo degenera nel valore costante pari a zero.

Abbiamo:

$X = x \rightarrow$ x insuccessi prima del primo successo durante le ripetizioni indipendenti di un esperimento di Bernoulli

$X \sim G(p) \rightarrow$ p indica la probabilità di successo dell'esperimento bernoulliano

Funzione di massa di probabilità

$p_X(i) = p(1-p)^i I_{N \cup \{0\}}(i)$

Vediamo che il supporto è infinito, in quanto non so mai quando si verifica il primo successo. La variabili i indica gli induccessi.

$\sum_{i=0}^{+\infty} p_X(i) = \sum_{i=0}^{+\infty} p(1-p)^i = p\sum_{i=0}^{+\infty}(1-p)^i = p\frac{1}{1-(1-p)} = 1$

Vediamo ora un breve lemma che ci concederà di calcolare più agevolmente il valore atteso.

Per ogni $\alpha \in (-1, 1)$

$\sum_{i=0}^{+\infty} i\alpha^i = \frac{\alpha}{(1-\alpha)^2}$

Dimostrazione.

$\sum_{i=0}^{+\infty} i\alpha^i = \alpha\sum_{i=0}^{+\infty} i\alpha^{i-1}\ \alpha\sum_{i=0}^{+\infty}\frac{d}{d\alpha}\alpha^i = \alpha\frac{d}{d\alpha}\sum_{i=0}^{+\infty}\alpha^i = \alpha\frac{d}{d\alpha}\frac{1}{1-\alpha} = \alpha\frac{1}{(1-\alpha)^2}$

Valore atteso

$E(X) = \sum_{i=0}^{+\infty} ip_X(i) = \sum_{i=0}^{+\infty} ip(1-p)^i = p\sum_{i=0}^{+\infty} i(1-p)^i$

e quindi per il lemma appena dimostrato si ha, ponendo $\alpha = 1-p$,

$E(X) = p\frac{1-p}{p^2} = \frac{1-p}{p}$

Varianza

$E(X^2) = \sum_{i=0}^{+\infty} i^2 p_X(i) = \sum_{i=0}^{+\infty} i^2 p(1-p)^i = p(1-p)sum_{i=0}^{+\infty} i^2(1-p)^{i-1} = p(1-p)\sum_{i=0}^{+\infty}\frac{d}{dp}(-i(1-p)^i) = -p(1-p)\frac{d}{dp}\sum_{i=0}^{+\infty} i(1-p)^i$

Applicando il lemma visto precedentemente, abbiamo:

$E(X^2) = -p(1-p)\frac{d}{dp}\frac{1-p}{p^2} = p(1-p)\frac{p^2+1p(1-p)}{p^4} = \frac{(1-p)(2-p)}{p^2}$

Possiamo ora calcolare la varianza come:

$Var(X) = E(X^2) - E(X)^2 = \frac{(1-p)(2-p)}{p^2} - \frac{(1-p)^2}{p^2} = \frac{1-p}{p^2}$

Funzione di ripartizione

Concentriamoci ora su un evento $X > n$ con $n \in N$ generico. Qual è la formalizzazione di tale probabilità?

$P(X > n) = \sum_{i=n+1}^{+\infty} p_X(i) = \sum_{i=n+1}^{+\infty}(1-p)^i p = p(1-p)^{n+1}\sum_{i=n+1}^{+\infty}(1-p)^{i-(n+1)} = p(1-p)^{n+1}\sum_{j=0}^{+\infty}(1-p)^j = p(1-p)^{n+1}\frac{1}{1-(1-p)} = (1-p)^{n+1}$

Pertanto, fissato $n \in N$ si avrà

$F_X(n) = P(X \leq n) = 1 - P(X > n) = 1 - (1-p)^{n+1}$

Fissato invece un generico $x \in R^+$ e indicato con $\lfloor x \rfloor$ l'intero ottenuto troncando x (o, equivalentemente, arrotondandolo per difetto), l'evento $X \leq x$ equivarrà a $X \leq \lfloor x \rfloor$. Otteniamo quindi:

$F_X(x) = (1-(1-p)^{\lfloor x \rfloor + 1})I_{[0,+\infty]}(x)$

Si noti infine che $P(X \geq x) = P(X \geq \lfloor x \rfloor) = P(X > \lfloor x \rfloor - 1) = (1-p)^{\lfloor x \rfloor}$, e quindi

$P(X \geq x + y | X \geq x) = \frac{P(X \geq x+y, X \geq x)}{P(X \geq x)} = \frac{P(X \geq x+y)}{P(X \geq x)} = \frac{(1-p)^{\lfloor x \rfloor + \lfloor y \rfloor}}{(1-p)^{\lfloor x \rfloor}} = (1-p)^{\lfloor y \rfloor} = P(X \geq y)$

Questa proprietà prende il nome di *assenza di memoria*. Essa indica che durante la ripetizione dell'esperimento bernoulliano, il fatto che sia avvenuto un numero n (anche elevato) di insuccessi

consecutivi non permette di dire alcunché sul numero di successivi insuccessi prima che si verifichi il primo successo. In altre parole, non c'è nessuna differenza, da un punto di vista probabilistico, dalla ripetizione degli esperimenti che vanno dal $n+1$-esimo in poi e dal ricominciare da capo la ripetizione.

Modello di Poisson

Consideriamo:

$\lambda \in R^+$

$X \sim P(\lambda)$

Funzione di massa di probabilità

$p_X(i) = P(X = i) = e^{-\lambda}\frac{\lambda^i}{i!}I_{N\cup\{0\}}(i)$

Vediamo di confermarne la correttezza:

1. Non negatività: è non negativa in quanto operante con tutti valori maggiori di 0
2. La somma dei valori assoluti della funzione è uguale a 1. Dimostriamolo:

$\sum_{i=0}^{+\infty} p_X(i) = \sum_{i=0}^{+\infty} e^{-\lambda}\frac{\lambda^i}{i!} = e^{-\lambda}\sum_{i=0}^{+\infty}\frac{\lambda^i}{i!} =$ $\rightarrow$ utilizzando Taylor $\rightarrow$ $= e^{-\lambda}e^{\lambda} = 1$

Valore atteso

$E(X) = \sum_{i=0}^{+\infty} ip_X(i) = \sum_{i=0}^{+\infty} ie^{-\lambda}\frac{\lambda^i}{i!} = \sum_{i=1}^{+\infty} ie^{-\lambda}\frac{\lambda^i}{i!} = e^{-\lambda}\sum_{i=1}^{+\infty}\frac{\lambda^i}{(i-1)!} = e^{-\lambda}\lambda\sum_{i=1}^{+\infty}\frac{\lambda^{i-1}}{(i-1)!} = e^{-\lambda}\lambda\sum_{i=0}^{+\infty}\frac{\lambda^i}{i!} = \lambda e^{-\lambda}e^{\lambda} = \lambda$

Varianza

$E(X^2) = \lambda + \lambda^2$ Dimostrazione lasciata al lettore

$Var(X) = E(X^2) - E(X)^2 = \lambda + \lambda^2 - \lambda^2 = \lambda = E(X)$

Quindi valore atteso e varianza coincidono. Consideriamo:

$X \sim B(n, p)$

Supponiamo di considerare tante variabili aleatorie binomiali, dove il parametro n è di volta in volta più grande ($n \rightarrow +\infty$). Vogliamo che il parametro p diminuisca di conseguenza, in modo da mantenere costante il prodotto np. Ossia, fissato un certo λ deve valere $np = \lambda \rightarrow p = \frac{\lambda}{n}$

Quindi possiamo interpretare la distribuzione di Poisson come la distribuzione limite che si ottiene considerando la distribuzione binomiale e facendo tenedere a $+\infty$ il numero di prove (diminuendo p in modo che $np = \lambda$ fissata).

Quindi, una variabile aleatoria con distribuzione di Poisson può essere usata come approssimazione di una variabile aleatoria binomale, purché siano n abbastanza grande e p abbastanza piccola. Dunque, formalizzando, dato:

$X \sim B(n, p)$ con n grande e p piccolo

Indichiamo con:

$X \dot{\sim} P(\lambda)$

l'approssimazione di X con distribuzione di Poisson, dove $\lambda = np$

Vediamo ora la proprietà di riproducibilità della distribuzione di Poisson. Dati:

1. $X_1 \sim P(\lambda_1)$
2. $X_2 \sim P(\lambda_2)$

Abbiamo:

$X_1 + X_2 \sim P(\lambda_1 + \lambda_2)$

Quindi, la distribuzione della somma fa parte della stessa famiglia delle distribuzioni originarie.

Modello ipergeometrico

Sappiamo che la distribuzione binomiale è la descrizione formale della ripetizione di un esperimento bernoulliano in cui vengono calcolati il numero di successi. Immaginiamo ora di avere una ripetizione di un esperimento dove vengono fatte una serie di estrazioni: in questo caso il numero di successi è ancora descrivibile tramite una trasformazione binomiale? Dipende dalla reimmissione: se è presente posso rappresentare l'esperimento con una binomiale, altrimenti devo utilizzare la distribuzione ipergeometrica. Consideriamo il caso in cui abbiamo:

- N oggetti corretti
- M oggetti errati
- $N + M$ oggetti totali

X = numero di oggetti corretti estratti senza reimmissione

Allora avremo:

$P(X = i) = \frac{\binom{N}{i}\binom{M}{n-i}}{\binom{N+M}{n}}$

Definisco la variabile aleatoria bernoulliana X_i che può assumere valori:

$X_i = 1$ se l'i-esima estrazione è un successo $X_i = 0$ altrimenti

Valore atteso

$E(X_i) = \frac{N}{N+M} := p$

$X = \sum_{i=1}^{n} X_i \rightarrow$ numero di successi su n estrazioni

Quindi:

$E(X) = E(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} E(X_i) = n\frac{N}{N+M} := np$

Se interpreto $\frac{N}{N+M}$ come p sembrerebbe non esserci differenza fra distribuzione binomiale e ipergeometrica. Tuttavia, se col valore atteso sembrano non esserci differenze, il discorso non è il medesimo con la varianza.

Varianza

$Var(X_i) = \frac{N}{N+M}(1 - \frac{N}{N+M}) = \frac{NM}{(N+M)^2}$

Covarianza

$Cov(X_i, X_j) = E(X_i X_j) - E(X_i)E(X_j)$

Il prodotto di due variabili bernoulliane è anch'esso una variabile bernoulliane. La covarianza, nel caso della binomiale, sarebbe stata pari a 0 in quanto, in quel caso, vi è indipendenza fra X_i e X_j. Ora non posso dire più la stessa cosa.

$Cov(X_i, X_j) = \frac{-NM}{(N+M)^2(N+M-1)}$

La dimostrazione è lasciata al lettore.

Quindi posso vedere la varianza come:

$Var(X) = Var(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} Var(X_i) + \sum_{i \neq j} Cov(X_i, X_j)$

Da cui, attraverso calcoli, ottengo:

$Var(X) = np(1-p)\left(1 - \frac{n-1}{N+M-1}\right)$

Il termine $\left(1 - \frac{n-1}{N+M-1}\right)$ era mancante nel caso binomiale. Esso è introdotto a causa della non indipendenza. Una cosa interessante da osservare è che all'aumentare del numero N di oggetti da estrarre e mantenendo fissato il numero di estrazioni da poter fare, $N + M \to \infty$. Ma se ciò accade allora:

$1 - \frac{n-1}{N+M-1} \to 1$

Quindi sembrerebbe che all'aumentare di $N + M$ la distribuzione ipergeometrica possa essere approssimata alla distribuzione binomiale.

1.11 Modellizzazione continua

Modello uniforme continuo

Consideriamo:

$a, b \in R \qquad a < b$

$X \sim U([a, b])$

Funzione di densità

Un generico valore appartenente all'intervallo $[a, b]$ ha la stessa densità. Definisco:

$f_U(x) = kI_{[a,b]}(x)$

Sappiamo che:

$\int_a^b f_U(x)dx = 1 \to f_U(x) = \frac{1}{b-a} I_{[a,b]}(x)$

Quindi, sostituendo all'integrale:

$\int_a^b \frac{1}{b-a} dx = \frac{1}{b-a} \int_a^b dx = \frac{1}{b-a}[x]_a^b = \frac{b-a}{b-a} = 1$

Funzione di ripartizione

$F_U(x) = P(X \leq x) = \int_{-\infty}^{x} f_U(n)dn$

Tale funzione può assumere tre diversi valori:

- $0 \qquad$ se $x < a$
- $\int_a^x \frac{1}{b-a} dn \qquad$ se $x \in [a, b]$

- $1 \qquad$ se $x > b$

$\int_a^x \frac{1}{b-a} dn = \frac{1}{b-a}[n]_a^x = \frac{x-a}{b-a}$

Quindi i valori assumibili sono:

- $0 \qquad$ se $x < a$
- $\frac{x-a}{b-a} \qquad$ se $x \in [a, b]$
- $1 \qquad$ se $x > b$

Quindi possiamo formalizzare dicendo:

$F_U(x) = \frac{x-a}{b-a} I_{[a,b]}(x) + I_{(b,+\infty)}(x)$

Valore atteso

$E(X) = \int_{-\infty}^{+\infty} x f_U(x) dx = \int_a^b x \frac{1}{b-a} dx = \frac{a+b}{2}$

La dimostrazione dei passi intermedi è lasciata al lettore.

Varianza

$E(X^2) = \int_{-\infty}^{+\infty} x^2 f_U(x) dx = \int_a^b x^2 \frac{1}{b-a} dx = \frac{a^2+ab+b^2}{3}$

La dimostrazione dei passi intermedi è lasciata al lettore.

$Var(X) = E(X^2) + E(X)^2 = \frac{a^2-2ab+b^2}{12} = \frac{(b-a)^2}{12}$

La dimostrazione dei passi intermedi è lasciata al lettore.

Quindi la varianza di una variabile aleatoria continua uniforme dipende dal quadrato della lunghezza di un intervallo.

Modello esponenziale

Tipicamente la distribuzione esponenziale riguarda la distribuzione della quantità di tempo prima che si verifichi un particolare evento.

$X \sim E(\lambda) \to \lambda \in R^+$

$f_X(x) = \lambda e^{-\lambda x} I_{R^+}(x)$

Funzione di densità

$\int_{-\infty}^{+\infty} f_X(x) dx = \int_0^{+\infty} \lambda e^{-\lambda x} dx$

Applico la sostituzione: $\lambda x = z$

$\int_0^{+\infty} e^{-z} dz = [-e^{-z}]_0^{+\infty} = -0 + 1 = 1$

Funzione di ripartizione

$F_X(x) = P(X \leq x) = \int_{-\infty}^x f_X(n) dn = \int_0^x \lambda e^{-\lambda n} dn$

Applico la sostituzione: $\lambda n = v$

$\int_0^{\lambda x} e^{-v} dv = [-e^{-v}]_0^{-\lambda x} = -e^{-\lambda x} + 1$

$\to F_X(x) = (1 - e^{-\lambda x}) I_{R^+}(x)$

Valore atteso

$E(X) = \int_0^{+\infty} x\lambda e^{-\lambda x} dx = [-xe^{-\lambda x}]_0^{+\infty} + \int_0^{+\infty} e^{-\lambda x} dx =$

$= \frac{1}{\lambda} \int_0^{+\infty} \lambda e^{-\lambda x} dx = \frac{1}{\lambda} \int_0^{+\infty} f_X(x) dx = \frac{1}{\lambda}$

All'aumentare di λ, la curva di f_x inzia più in alto, ma decresce più velocemente, poiché l'area sottesa è pari a 1. Dunque i valori avranno maggiore densità vicino allo 0, facendo sì che anche il valore atteso tenda ad essere più piccolo (tendenzialmente, vicino allo 0 anch'esso).

Varianza

$E(X^2) = \int_0^{+\infty} x^2 \lambda e^{-\lambda x} dx = [-x^2 e^{-\lambda x}]_0^{+\infty} + \int_0^{+\infty} 2xe^{-\lambda x} dx =$

$\frac{2}{\lambda} \int_0^{+\infty} x\lambda e^{-\lambda x} dx = \frac{2}{\lambda} E(X) = \frac{2}{\lambda} \frac{1}{\lambda} = \frac{2}{\lambda^2}$

$Var(X) = E(X^2) - E(X)^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}$

All'aumentare di λ, i valori tendono ad essere vicini a 0, ovvero tendono ad appartenere ad un insieme più piccolo. Avendo una varianza più piccola, ne consegue che avremo anche una minore deviazione standard.

Assenza di memoria

La proprietà di decadere della funzione di densità coincide con l'assenza di memoria. Vediamo di formalizzare:

$P(X > s + t | X > s) = P(X > t)$

E' indifferente utilizzare il $>$ o il $\geq$. Tale uguaglianza vale $\forall s, t \geq 0$

$\frac{P(X>s+t|X>s)}{P(X>s)} = P(X > t)$

$\frac{P(X>s+t)}{P(X>s)} = P(X > t)$

$P(X > s + t) = P(X > t)P(X > s)$

$F_X(x) = 1 - e^{-\lambda x} = P(X \leq x)$

$P(X > x) = 1 - P(X \leq x) = 1 - F_X(x) = e^{-\lambda x}$

$e^{-\lambda(s+t)} = e^{-\lambda t} e^{-\lambda s}$

Ho ottenuto una identità, dunque vale l'assenza di memoria. L'unica distribuzione continua con assenza di memoria è quella esponenziale.

Proprietà 1

Dati:

$X_1, ..., X_n$ indipendenti

$\forall i \quad X_i \sim E(\lambda_i)$

$Y = min_i X_i \to Y \sim E(\sum_{i=1} nX_i) = E(\lambda)$

La dimostrazione è lasciata al lettore.

Proprietà 2

Se considero:

$X \sim E(\lambda)$ e $Y = cX$ con $c > 0$, allora ho:

$P(Y \leq x) = P(cX \leq x) = ... = F_Y(x)$

Quindi, se X è una variabile aleatoria distribuita esponenzialmente, anche Y lo sarà.

$X \sim E(\lambda), c > 0 \rightarrow Y = cX \sim E(\frac{\lambda}{c})$

Modello normale o gaussiano

Consideriamo un esperimento di carattere binomiale, ripentendolo di continuo e aumentando il numero di ripetizioni di volta in volta fino all'infinito. Immaginiamo di disegnare i grafico a bastoncini della funzione di massa di probabilità della variabile binomiale, infittendolo fino al punto da avere un andamento continuo. Arriveremo ad avere un grafico a campana, dove ogni punto rappresenta la vetta di un bastoncino, andando a creare una curva continua di punti. La distribuzione gaussiana è la distribuzione a cui tende la binomiale quando il paramentro n tende a $+\infty$

$n \in R \qquad \sigma \in R^+ \qquad\qquad X \sim N(\mu, \sigma^2)$

Dove μ rappresenta il valore atteso e σ^2 la varianza.

Funzione di densità

Il dominio della distribuzione normale è R. La sua funzione di densità sarà:

$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

Il grafico di tale distribuzione è simmetrico rispetto a μ, che ne rappresenta anche il massimo globale.

La distribuzione normale sembra essere una buona approssimazione della distribuzione binomiale. Tuttavia, sovrapponendo il grafico della funzione di densità della distribuzione normale con quello della funzione di massa di probabilità della distribuzione binomiale ci rendiamo conto che non sono simili: uno ha altezza massima pari a 1, l'altro integra ad 1.

Verifichiamo quindi che $f_X(x)$ sia una funzione di densità, immaginando $\mu = 0$ e $\sigma = 1$:

$\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{x^2}{2}} dx$

Consideriamo solo l'integrale e identifichiamolo con I:

$I^2 = \int_{-\infty}^{+\infty} e^{-\frac{x^2}{2}} dx \int_{-\infty}^{+\infty} e^{-\frac{y^2}{2}} dy = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} e^{-\frac{x^2}{2}} e^{-\frac{y^2}{2}} dxdy = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} e^{-\frac{x^2+y^2}{2}} dxdy = 2\pi$

Quindi:

$I = \sqrt{2\pi}$

Torniamo infinite al passo prima dell'analisi dell'integrale e sostituiamo il valore I:

$\frac{1}{\sqrt{2\pi}} \sqrt{2\pi} = 1$

Ciò accade quando $\mu = 0$ e $\sigma = 1$. In realtà, se non avessi considerato queste condizioni, nell'integrale ove ho sostituito y a x avrei dovuto considerare $y = \frac{x-\mu}{\sigma}$. Ma non sarebbe cambiato nulla, poiché $f_X(x)$ integra ad 1 indipendentemente da μ e σ.

Immaginiamo di ridurre σ: il punto di massimo della funzione sarà più alto, poiché la funzione deve integrare a 1. Ciò lo posso fare per σ qualsiasi, dunque esisterà per forza un σ per cui il punto

di massimo supererà uno e, quindi, ergo che i due grafici, per quanto coerenti fra di loro, non si sovrappongono. Si può dimostrare che esiste un fattore moltiplicativo costante per passare da una funzione all'altra.

Valore atteso

$E(X) = \mu$

Varianza

$Var(X) = \sigma^2$

Vediamo di riprendere alcuni vecchi concetti espressi all'inizio del documento e applicandoli alla modellizzazione.

1. Moda: valore per cui la funzione di massa di probabilità (di una distribuzione discreta) o la funzione di densità (di una distribuzione continua) hanno il massimo. Nel caso di una distribuzione normale, tale massimo sarà μ
2. Mediana: si ricollega alla funzione di ripartizione. Nel caso continuo è necessario spaccare la curva di densità in modo tale che l'area di sinistra sia uguale all'area di destra. Nel caso della distribuzione normale, la mediana sulla quale applicare il taglio è μ. Da ciò intuiamo che μ è media, moda e mediana
3. Quantili/Quartili: μ permette di arricchire il QQ-Plot. Lo stesso discorso può essere fatto per qualsiasi distribuzione, che essa sia continua o discreta

Quindi, invece di confrontare due campioni per la stesura del QQ-Plot, possono confrontare un campione e la distribuzione normale dei parametri dati. Tale confronto è così frequente che esiste una variante del QQ-Plot che confronta direttamente i dati con la distribuzione normale. E' necessario notare che a seconda dell'implementazione, la retta risultante può nono essere la bisettrice del primo e terzo quadrante.

Riproducibilità

Presi:

$X_1 \sim N(\mu_1, \sigma_1^2)$

$X_2 \sim N(\mu_2, \sigma_2^2)$

Indipendenti fra diloro, abbiamo:

$X_1 + X_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$

E' interessante osservare la linearità di una distribuzione normale:

$X \sim N(\mu, \sigma^2) \rightarrow aX + b \sim N(a\mu + b, a^2\sigma^2) \qquad \forall a, b \in R, a \neq 0$

Funzione di ripartizione

$F_X(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-\mu)^2}{2\sigma^2}} dy = \int_{-\infty}^{\frac{x-\mu}{\sigma}} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz =$

Se la distribuzione è del tipo $N(0, 1)$ possiamo scrivere:

$= \Phi\left(\frac{x-\mu}{\sigma}\right)$

Tale funzione è chiamata *normale standard* e mi permette di calcolare la funzione di ripartizione di una distribuzione normale qualsiasi:

$P(a \leq X \leq b) = F_X(b) - F_X(a) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$

Quindi abbiamo il passaggio:

$X \to Z = \frac{X_\mu}{\sigma}$

Con:

$X \sim N(\mu, \sigma^2)$

$Z \sim N(0, 1)$

Questo processo, chiamato *standardizzazione* gode della seguente proprietà:

$F_X(x) = P(X \leq x) = P\left(\frac{X-\mu}{\sigma} \leq \frac{x-\mu}{\sigma}\right) = P\left(Z \leq \frac{x-\mu}{\sigma}\right) = \Phi\left(\frac{x-\mu}{\sigma}\right)$

La funzione di densità, una volta applicata la standardizzazione, viene traslata per far sì che il massimo coincida con il punto 1 e che il grafico sia simmetrico rispesso all'asse delle ordinate. L'area da considerare sarà quella che va da $-x$ a x.

Proprietà

$\Phi(-x) = \int_x^{+\infty} \phi(x)dx = P(Z > x) = 1 - P(Z \leq x) = 1 - \Phi(x)$

Teorema centrale del limite

Abbiamo già parlato del fatto che si può ottenere una distribuzione normale come il limite della distribuzione binomiale con il parametro n che tende a più infinito. Con il teorema centrale del limite cerchiamo di dare un'altra interpretazione, dimostrata da Alan Turing.

$X_1, X_2, ..., X_n$

Variabili aleatorie indipendenti e identicamente distribuite (IID), ossia hanno la stessa identica distribuzione con stessi valori di parametri. Equivarrebbe a dire che tali variabili sono distribuite identicamente come X. Abbiamo:

$\sum_{i=1}^n X_i \dot{\sim} N(n\mu, n\sigma^2)$

Il simbolo "$\dot{\sim}$" è da leggersi "approssivamente distribuito come". Inoltre, abbiamo che:

$E(X) = \mu$

$Var(X) = \sigma^2$

L'approssimazione è tanto più buona tanto più è grande n. Applicando la standardizzazione, abbiamo:

$\frac{\sum_{i=1}^n X_i - n\mu}{\sigma\sqrt{n}} \dot{\sim} N(0, 1)$

Quindi:

$P\left(\frac{\sum_{i=1}^n X_i - n\mu}{\sigma\sqrt{n}} \leq x\right) \approx \Phi(x)$

Vediamo ora un esempio di applicazione di standardizzazione nel caso di esperimenti bernoulliani. Supponiamo di avere n variabili aleatorie:

$X_1, ..., X_n \qquad \forall i = 1, ..., n \to X_i \sim B(p)$

$\forall i \qquad E(X_i) = p \quad Var(X_i) = p(1-p)$

Se consideriamo:

$X = \sum_{i=1}^{n} X_i$

avremo che il valore atteso e la varianza verranno modificati di conseguenza:

$E(X) = np$

$Var(X) = np(1-p)$

Quindi avremo:

$X \dot{\sim} N(np, np(1-p))$

Che standardizzato diviene:

$\frac{X-np}{\sqrt{np(1-p)}} \dot{\sim} N(0,1)$

2 Statistica Inferenziale

Inferenza: processo di induzione a partire da osservazione faccio delle ipotesi. Ipotizzo che una particolare distribuzione descriva bene le osservazioni sul campione. Non conosco la distribuzione a priori

Deduzione: da verità generale a una sua istanziazione

Induzione: come metodo scientiico, da manifestazioni si cerca di derivare verità generale. E' il processo inverso rispetto alla deduzione. Si cerca di passare da verità specifiche a verità generali. Ovviamente questo passaggio implica dei rischi

La statistica inferenziale mi permette di capire quale distribuzione descrive bene le osservazioni. Essa è caratterizzata da due metodologie:

1. Non parametrica: senza alcuna conoscenza (non trattata nel corso)
   - Partendo da dati che osservo e sapendo che sono distribuiti secondo una certa famiglia di distribuzioni, ne ricvo i parametri che consentono una buona approssimazione di tali dati tramite quella stessa distribuzione. Partendo da un insieme di osservazioni, ipotizzo una distribuzione dei dati (approssimazione di una variabile aleatoria) ricavandone un'approssimazione dei parametri della distribuzione
2. Parametrica: parto da una conoscenza di base, conoscendo la famiglia di distribuzione ma se senza conoscerne i parametri (o alcuni di essi)

Campione: successione $X_1, X_2, ..., X_n$ di variabili aleatorie IID. n indica la grandezza del campione. Se vi è indipendenza fra variabili aleatorie parliamo di campione casuale.

Statistica/Stimatore: algoritmo applicato a un campione il cui output serve come approssimazione dei parametri della distribuzione non noti. E' una variabile aleatoria il cui valore è determinato dai dati del campione. Deve dipendere esclusivamente dal campione (casuale) non dai parametri ignoti. Un esempio di stimatore è la media campionaria, che si comporta da statistica per il valore atteso:

$E(X) = \mu$

$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} X_i$

Abbiamo che:

$\overline{x} \approx E(X) = \mu$

Vediamo di dimostrarlo:

$E(\overline{X}) = E(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n}\sum_{i=1}^{n} \mu = \mu$

La media campionaria varia ad ogni osservazione, ma tende ad oscillare attorno ad un valore centrale: μ Dunque possiamo dire che ne rappresenta una buona approssimazione. Vediamo come si comporta la media campionaria con la varianza:

$Var(\overline{X}) = Var(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n^2} Var(\sum_{i=1}^{n} X_i) = \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$

All'aumentare di taglia del campione, diminuisce l'oscillazione dal valore centrale (ossia la varianza). Quindi, più è grande il campione e minore è l'errore che è commesso nelle osservazioni.

In generale:

Popolazione: descritta da una variabile aleatoria X tale che $X \sim F_X(\bullet;\theta)$. Con tale notazione descriviamo implicitamente una distribuzione che vuole essere quanto più generica possibile. θ rappresenta un parametro ignoto

Scopo: trovare un valore sensato del parametro ignoto θ a partire da una statistica

Statistica:

$t(X_1, ..., X_n) = T$

Numero potenzialmente variabile di argomenti. Se voglio calcolare il valore preciso uso i valori delle osservazioni invece di variabili aleatorie (ossia i valori campionari $x_1, ..., x_n$).

$t(x_1, ..., x_n) \approx \theta$ è un'approssimazione di θ.

Oppure posso considerare:

$t(x_1, ..., x_n) = \dot{\theta}$ dove $\dot{\theta}$ è una stima di θ, quindi:

$\dot{\theta} \approx \theta$

N.B. $x_1, x_2, ..., x_n$ sono i valori osservati, la realizzazione campionaria (e.g. dataset). Non necessariamente però sono interessato a stimare θ.

In generale, una popolazione può essere descritta da una variabile aleatoria X tale che $X \sim F_x(\bullet;\theta)$. Il nostro interesse è stimare:

$\tau(\theta)$ con caso particolare del tipo: $\tau(\theta) = \theta$

Allora:

$t(x_1, ..., x_n) \approx \tau(\theta)$

ossia:

$t(x_1, ..., x_n) = \dot{\tau}$ dove $\dot{\tau} \approx \tau(\theta)$

Quando succede che:

$E(t(X_1, ..., X_n)) = \tau(\theta)$

Diciamo che:

$t(X_1, ..., X_n)$ è non deviato rispetto a $\tau(\theta)$ (oppure che è corretto, senza deviazione, unbiased, **non distorto**)

N.B. Se sono interessato a stimare il valore atteso di una popolazione, posso usare la media campionaria come stima del valore atteso, indipendentemente dalla distribuzione.

Vediamo come funziona la varianza campionaria come stimatore.

$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 \to S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$

$E(S^2) = \sigma^2$

Quindi S^2 è uno stimatore non deviato per σ^2 (varianza della popolazione). Reciprocamente abbiamo:

T deviato rispetto a $\tau(\theta)$

$E(T) \neq \tau(\theta) \to E(T) - \tau(\theta)$

Immaginiamo ora di avere una popolazione $X \sim U([0, \theta]) \quad \theta > 0$

$\tau(\theta) = \theta$

Se avessimo $X_1, ..., X_n$ facenti parte del campione, come posso stimare θ?

$t(X_1, ..., X_n) = max_{1 \leq i \leq n} X_i$

E' deviato o no? Dobbiamo verificare che:

$E(T) = E(maxX_i) = \theta?$

$E(T) = \int_0^\theta t f_T(t) dt$

$F_T(t) = P(T \leq t) = P(max_{1 \leq i \leq n} X_i \leq t) = P(\forall i X_i \leq t) = P(\bigcap_{i=1}^{n} X_i \leq t) = \Pi_{i=1}^{n} P(X_i \leq t) = \Pi_{i=1}^{n} F_{X_i}(t) = F_X(t)^n = \left(\frac{1}{\theta}\right)^n$

So che $F_T(t) = \left(\frac{t}{\theta}\right)^n$

$f_T(t) = \frac{dF_T(t)}{dt} = n\left(\frac{t}{\theta}\right)^{n-1}\frac{1}{\theta} = \frac{nt^{n-1}}{\theta^n}$

Dunque:

$E(T) = \int_0^\theta t f_T(t) dt = \int_0^\theta t \frac{nt^{n-1}}{\theta^n} dt = \frac{n}{\theta^n}\int_0^\theta t^n dt = \frac{n}{\theta^n}[\frac{t^{n+1}}{n+1}]_0^\theta = \frac{n}{\theta^n}\frac{\theta^{n+1}}{n+1} = \frac{n}{n+1}\theta$

Lo stimatore è deviato, sebbene per n grande la frazione tende a 1. N.B. Diventa non deviato se lo moltiplico per $\frac{n+1}{n}$

Quindi, ricapitolando:

$T = \overline{X} \qquad E(T) = E(X) = \frac{\theta}{2}$

Idea: uso $2\overline{X}$

$E(2\overline{X}) = 2E(\overline{X}) = 2\frac{\theta}{2} = \theta$

Quindi uno stimatore interessante è:

$T' = \frac{2}{n}\sum_{i=1}^{n} X_i$

N.B. Per un dato parametro lo stimatore non deviato non è unico

Introduciamo un concetto di dispersione, che fornisce un'idea sulla precisione dello stimatore.

$|T - \tau(\theta)|$ è l'errore commesso

$E(|T-\tau(\theta)|$ è difficile da calcolare, l'abbiamo già visto con la varianza. Infatti utilizziamo gli scarti quadratici:

$E((T - \tau(\theta))^2) = MSE_{\tau(\theta)}(T)$

Tale uguaglianza identifica lo scarto quadratico medio. Esso offre un'idea sull'ordine di grandezza dell'errore commesso nello stimare $\tau(\theta)$ con T.

$MSE_{\tau(\theta)}(T) = E((T - E(T) + E(T) - \tau(\theta))^2) = E((T - E(T))^2 + 2(T - E(T))(E(T) - \tau(\theta)) + (E(T) - \tau(\theta))^2) = E((T - E(T))^2 + E(2(T - E(T))(E(T) - \tau(\theta))) + (E(T) - \tau(\theta))^2) = Var(T) + 2(E(T) - \tau(\theta))E(T - E(T)) + (E(T) - \tau(\theta))^2 = Var(T) + (E(T) - \tau(\theta))^2$

$MSE_{\tau(\theta)}(T) = Var(T) + b_{\tau(\theta)}(T)^2$

T non distorto rispetto a $\tau(\theta)$ se:

$b_{\tau(\theta)}(T) = 0 \rightarrow MSE_{\tau(\theta)}(T) = Var(T)$

Osservazione:

$lim_{n\rightarrow+\infty} MSE_\theta(T_n) = 0$

Consistenza in media quadratica. Immaginiamo di ripetere il campionamento aumentando di volta in volta la taglia del campione. Quindi l'errore diminuisce all'aumentare della taglia del campione. Se siamo confindenti di aver preso un numero sufficientemente grande di dati, allora l'errore sarà molto basso, dunque la statistica approssimerà bene il parametro θ (o $\tau(\theta)$)

Consistenza debole in media quadratica

Stimatore T_n di $\tau(\theta)$. T_n non è uno stimatore esatto, ma non si distoglie troppo da $\tau(\theta)$

$\tau(\theta) - \epsilon < T_n < \tau(\theta) + \epsilon$

Con $\epsilon > 0$

$lim_{n\rightarrow+\infty} P(\tau(\theta) - \epsilon < T_n < \tau(\theta) + \epsilon) = 1 \qquad \forall \epsilon > 0$

Allora vale la consistenza debole. Una cosa da notare è che la consistenza implica la consistenza debole. Dimostriamolo:

$P(-\epsilon < T_n - \tau(\theta) < \epsilon) = P(|T_n - \tau(\theta)| < \epsilon) \rightarrow P((T_n - \tau(\theta))^2 < \epsilon^2)$

Applichiamo la disuguaglianza di Markov:

$P((T_n - \tau(\theta))^2 < \epsilon^2) \geq 1 - \frac{E((T_n - \tau(\theta))^2)}{\epsilon^2} = 1 - \frac{MSE_{\tau(\theta)}(T_n)}{\epsilon^2}$

$\implies lim_{n\rightarrow+\infty} P(-\epsilon < T_n - \tau(\theta) < \epsilon) \geq 1 - lim_{n\rightarrow+\infty} \frac{MSE_{\tau(\theta)}(T_n)}{\epsilon^2} = 1$

Nel caso di consistenza, abbiamo che:

$lim_{n\rightarrow+\infty} MSE_{\tau(\theta)}(T) = 0$

quindi il limite è uguale a 0 indipendentemente da ϵ

Considerazioni sul valore atteso

Abbiamo visto che nel caso discreto e continuo, il valore atteso viene calcolato mediante due formule diverse:

- Discreto: $E(X) = \sum_i x_i P(X = x_i)$
- Continuo: $E(X) = \int_{-\infty}^{+\infty} x f_X(x) dx$

Ora immaginiamo che X sia una variabile aleastoria discreta su $N \cup \{0\}$:

$E(X) = \sum_{i=0}^{+\infty} iP(X = i) = 0P(X = 0) + 1P(X = 1) + 2P(X = 2) + 3P(X = 3) + ... =$

$= P(X = 1) + P(X = 2) + P(X = 2) + P(X = 3) + P(X = 3) + P(X = 3) + ...$

Possiamo vedere $P(X = 1) + P(X = 2) + P(X = 3)...$ come $P(X > 0)$, $P(X = 2) + P(X = 3) + ...$ come $P(X > 1)$, $P(X = 3) + ...$ come $P(X > 2)$ e così via. Quindi abbiamo:

$= P(X > 0) + P(X > 1) + P(X > 2) + ... = \sum_{i=0}^{+\infty} P(X > i) = \sum_{i=0}^{+\infty} (1 - P(X \leq i) = \sum_{i=0}^{+\infty} (1 - F_X(i))$

Si ottiene un risultato simile anche per una variabile aleatoria X su R^+:

$E(X) = \int_0^{+\infty} (1 - F_X(x)) dx = \frac{1}{\lambda}$

Posso applicare la distribuzione esponenziale per verificarlo:

$X \sim E(\lambda)$

$\int_0^{+\infty} e^{-\lambda x} dx = \frac{1}{\lambda} \int_0^{+\infty} e^{-\lambda x} \lambda dx = \frac{1}{\lambda} \int_0^{+\infty} e^{-y} dy = \frac{1}{\lambda} [-e^{-y}]_0^{+\infty} = \frac{1}{\lambda}(0 + 1) = \frac{1}{\lambda}$

Legge dei grandi numeri

$\overline{X}_n = \frac{1}{n} \sum X_i \dot{\sim} \frac{1}{n} N(n\mu, n\sigma^2) \sim N(\mu, \frac{\sigma^2}{n})$

All'aumentare della taglia del campione, la media campionaria è approssivamente distribuita come una normale, la cui varianza tende a 0. Non abbiamo più una quantità aleatoria, bensì una costante. Si può fare questo passaggio senza complicazioni? Si, attraverso la legge dei grandi numeri.

Versione forte

$P(lim_{n \to +\infty} \overline{X}_n = \mu) = 1$

L'uguaglianza fra variabile aleatoria e un valore specifico indica un evento. Su di esso posso ragionare in termini di probabilità.

Versione debole

Alternativamente possiamo osservare:

$|\overline{X}_n - \mu|$

Questa quantità è a sua volta una variabile aleatoria, e ci aspettiamo che sia tanto più piccola al crescere di μ e, quindi, più piccola di una certa quantità $\epsilon > 0$.

Fissata $\epsilon > 0$:

$lim_{n \to +\infty} P(|\overline{X}_n - \mu| > \epsilon) = 0 \qquad \forall \epsilon > 0$

Standardizzazione

Abbiamo enunciato che $\overline{X}_n \dot{\sim} N(\mu, \frac{\sigma^2}{n})$. Possiamo quindi applicare la standardizzazione:

$\frac{\overline{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \dot{\sim} N(0,1)$

Non conosco ne μ ne σ^2. Per μ non abbiamo problemi, per σ^2 applichiamo una sostituzione inserendo una sua stima.

La varianza campionaria è uno stimatore non deviato della varianza:

$E(s^2) = \sigma^2$

$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2$

$(n-1)s^2 = \sum_{i=1}^{n} (X_i - \overline{X})^2$

Abbiamo già osservato precedentemente che:

$\sum_{i=1}^{n} (x_i - \overline{x})^2 = \sum_{i=1}^{n} (x_i^2 - 2x_i\overline{x} + \overline{x}^2) = \sum_{i=1}^{n} (x_i^2 - 2\overline{x} \sum_{i=1}^{n} x_i + n\overline{x}^2) = \sum_{i=1}^{n} x_i^2 - 2n\overline{x}^2 + n\overline{x}^2$

$\implies \sum_{i=1}^{n} (x_i - \overline{x})^2 = \sum_{i=1}^{n} x_i^2 - n\overline{x}^2$

Inoltre sappiamo che:

- $Var(X) = E(X^2) - E(X)^2$ -$E(X^2) = Var(X) + E(X)^2 = \sigma^2 + \mu^2$ -$E(\overline{X}^2) = Var(\overline{X}) + E(\overline{X})^2 = \frac{\sigma^2}{n} + \mu^2$
- $E(\overline{X}) = E(X_i) = \mu$
- $Var(X) = \sigma^2, Var(\overline{X}) = \frac{\sigma^2}{n}$

Torniamo indietro, abbiamo analizzato la parte a destra dell'uguaglianza:

$(n-1)s^2 = \sum_{i=1}^{n} (X_i - \overline{X})^2$

Ora osserviamo:

$(n-1)s^2 = \sum_{i=1}^{n} (X_i - \overline{X})^2 = \sum_{i=1}^{n} X_i^2 - n\overline{X}^2$

$E((n-1)s^2) = E(\sum (X_i - \overline{X})^2 = E(\sum X_i^2 - n\overline{X}^2) = \sum E(X_i^2) - nE(\overline{X}^2) = nE(X^2) - nE(\overline{X}^2) = n(E(X^2) - E(\overline{X}^2)) = n(\sigma^2 + \mu^2 - \frac{\sigma^2}{n} - \mu^2) = n\sigma^2 - \sigma^2 = (n-1)\sigma^2$

$\implies E((n-1)s^2) = (n-1)\sigma^2 \to (n-1)E(s^2) = (n-1)\sigma^2 \to E(s^2) = \sigma^2$

La varianza campionaria è uno stimatore non deviato per la varianza di una popolazione.

Determinare la taglia del campione

- X = misurazione
- $E(X) = \mu$
- $Var(X) = \sigma^2$
- $Z \sim N(0,1)$

$P(|\overline{X} - \mu| \leq r) \geq 1 - \delta$

Vogliamo minimizzare r, δ:

$P(|\overline{X} - \mu| \leq r) = P\left(\frac{|\bar{(}X) - \mu|}{\frac{\sigma}{\sqrt{n}}} \leq \frac{r}{\frac{\sigma}{\sqrt{n}}}\right) = P\left(\left|\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}\right| \leq \frac{r}{\sigma}\sqrt{n}\right) \approx P(|Z| \leq \frac{r}{\sigma}\sqrt{n}) = P(-\frac{r}{\sigma}\sqrt{n} \leq Z \leq \frac{r}{\sigma}\sqrt{n}) = \Phi(\frac{r}{\sigma}\sqrt{n}) - \Phi(-\frac{r}{\sigma}\sqrt{n}) = 2\Phi(\frac{r}{\sigma}\sqrt{n}) - 1 \geq 1 - \delta \implies \Phi(\frac{r}{\sigma}\sqrt{n}) \geq 1 - \frac{\delta}{2}$

- n = taglia campione
- r = errore

- $1 - \delta$ = soglia di confidenza

1) Risolviamo rispetto ad n

$\frac{r}{\sigma}\sqrt{n} \geq \Phi^{-1}(1 - \frac{\delta}{2})$

$n \geq \frac{\sigma^2}{r^2}\Phi^{-1}(1 - \frac{\delta}{2})^2$

2) Risolviamo rispetto ad r

$r \geq \frac{\sigma}{\sqrt{n}}\Phi^{-}1(1 - \frac{\delta}{2})$

3) Risolviamo rispetto a δ

$\frac{\delta}{2} \geq 1 - \Phi(\frac{r}{\sigma}\sqrt{n})$

- Più grande è la varianza, maggiore sarà il numero di valori da osservare
- Più piccolo è r, minore sarà l'errore massimo ammissibile, quindi saranno necessari più valori osservati
- Più piccolo è δ maggiore sarà la soglia di confidenza, più sarà stringente la mia richiesta di rispettare l'errore, dunque serviranno maggiori valori osservati

E' da notare che se non conoscessimo σ^2, è possibile stimarlo usando come statistica la varianza campionaria. Così facendo, aumenta anche l'approssimazione.

Processo di Poisson

Siamo interessati ad eventi in un intervallo temporale (e.g. quante chiamate riceviamo in un giorno). Tuttavia, è possibile evitare di fissare a priori il tempo facendo:

- $[0, t)$ intervallo temporale
- $N(t)$ numero di eventi verificati in $[0, t)$

Quando si pensa a una famiglia di variabili aleatorie con parametro t legato al tempo, si parla di *processi stocastici*. Il processo di Poisson è un processo stocastico.

Ipotesi:

1) $t = 0$ è l'istante iniziale. Partire da 0 o da un altro valore è indifferente.

$N(0) = 0 \implies$ essendo che il processo inizia all'istante 0, non occorrono eventi se si verifica solo tale istante temporale

2) Indipendenza tra intervalli temporali disgiunti (rispetto al numero di eventi che vi occorrono)

3) La distribuzione del numero di eventi in un dato intervallo dipende solo dalla lunghezza dell'intervallo

4) $lim_{h\to 0}\frac{P(N(h)=1)}{h} = \lambda \implies$ se consideriamo intervalli di lunghezza breve, la probabilità che occorra esattamente un evento in tale intervallo è proporzionale alla lunghezza dell'intervallo stesso

5) $lim_{h\to 0}\frac{P(N(h)>1)}{h} = 0 \implies$ se consideriamo intervalli abbbastanza brevi, la probabilità che si verifichino due o più eventi è 0. Quindi, in genere, in un intervallo breve si verifica al più un evento

Enunciato

$N(t) \sim P(\lambda t)$

Dimostrazione

Consideriamo una linea temporale lunga da 0 a t. Suddividiamo l'intervallo $[0, t)$ in n (grande a piacere) intervalli equiampi. Consideriamo ora l'evento:

$N(t) = k$

$(N(t) = k) = A \cup B$

A e B sono insiemi disgiunti. Ognuna delle k occorrenze si trova in un sotto intervallo diverso di A. Tutti gli altri si trovano in B. Quindi, possiamo tradurre il tutto come:

$P(N(t) = k) = P(A) + P(B)$

Possiamo scegliere n abbastanza grande tale da azzerare $P(B)$. Supponiamo che i k eventi occorrano nei primi k intervalli. Abbiamo che:

$P(N(t) = k) = P(A) + P(B) = P(A) = \left(\frac{\lambda t}{n}\right)^k \left(1 - \frac{\lambda t}{n}\right)^{n-k} \binom{n}{k}$

Tutto ciò rappresenta una binomiale, quindi:

$N(t) \sim B(n, \frac{\lambda t}{n})$

Al crescere di n il rapporto $\frac{\lambda t}{n}$ decresce, quindi per $n \to +\infty$ abbiamo che:

$N(t) \sim P(\lambda t) \implies P(N(t) = k) = \frac{(\lambda t)^k}{k!} e^{-\lambda t}$

Ora supponiamo di evidenziare gli istanti di occorrenza degli eventi nella nostra linea temporale. Possiamo provare a determinare la distribuzione degli X_i, ossia la sequenza dei tempi di interarrivo. Concentriamoci sulla seguente proprietà:

$P(X_1 > t) = P(N(t) = 0) = e^{-\lambda t}$

$\implies F_{X_1}(t) = 1 - P(X_1 > t) = 1 - e^{-\lambda t} \implies X_1 \sim E(\lambda)$

Ora immaginiamo di conoscere X_1:

$P(X_2 > t | X_1 = s) = P(nessun - evento - in(s, s+t] | X_1 = s) = P(nessun - evento - in(s, s+t]) = P(N(t) = 0) = e^{-\lambda t}$

Quindi abbiamo che:

$F_{X_2}(t) = 1 - e^{-\lambda t} = F_{X_1}(t)$

Lo stesso discorso vale per tutti gli altri X_i, quindi:

$\forall X_i \sim E(\lambda)$

E quindi, deduciamo che gli X_i sono indipendenti.