Analisi di serie storiche

Luca Mari, versione 17.12.15

Contenuti

I principali concetti introdotti in questo capitolo

Serie storiche e casualità

Sia data una successione di coppie $\langle\langle x_i, y_i\rangle\rangle$, i=1,...,n, in cui x_i è una variabile indipendente e y_i una variabile dipendente, y_i=y_i(x_i); nel caso in cui la successione $\langle x_i\rangle$ descrive gli istanti di tempo in cui sono acquisiti gli elementi della successione $\langle x_i\rangle$, $\langle x_i\rangle$=$\langle t_i\rangle$, il campione $\langle\langle t_i, y_i\rangle\rangle$ si chiama *serie storica* (o, in certi contesti, semplicemente *segnale*).

Nell'ipotesi che gli intervalli $[t_i, t_{i+1}]$ abbiano ampiezza costante, $t_{i+1}-t_i=\Delta t$, il termine Δt è chiamato *periodo di campionamento* e il suo inverso $f_c=1/\Delta t$ *frequenza di campionamento*. In questo contesto, spesso per brevità gli elementi y_i sono chiamati direttamente "campioni", e quindi la frequenza di campionamento si misura in campioni all'unità di tempo, per esempio campioni al secondo, e in tal caso quindi hertz, Hz. Per esempio, una serie storica con $f_c=5$ Hz contiene 5 campioni per ogni secondo, acquisiti ogni 1/5 = 0,2 s.

Secondo l'ipotesi di periodo / frequenza di campionamento costante, e dato che l'unità di tempo rimane da stabilire e quindi ci si può sempre riportare al caso in cui $\Delta t=1$, una serie storica può essere dunque

descritta come un campione $\langle y_i \rangle$ il cui indice i descrive istanti di tempo, e in conseguenza è interpretabile come una funzione, che può essere quindi rappresentata non solo su un grafico a dispersione, come dovrebbe essere per un generico campione bivariato, ma anche su un usuale grafico a linee.
Le serie storiche relative a fenomeni reali, per esempio di tipo fisico o economico, sono caratterizzate da valori y_i la cui dipendenza dal tempo è raramente espressa da una funzione di cui è nota l'espressione analitica, cioè da una funzione f tale che $y_i=f(t_i)$. Ciò non implica, d'altra parte, che la successione $\langle y_i \rangle$ sia "completamente causale": in molti casi, infatti, si possono riconoscere in essa regolarità di qualche genere, a cui è sovrapposta una componente casuale, chiamata abitualmente "rumore": dell'individuazione di tali regolarità si occupa l'*analisi delle serie storiche* (in inglese: *time series analysis*, TSA).
Per esemplificare il senso per cui tali regolarità potrebbero manifestarsi, studiamo come una serie storica può essere generata. Nel caso più semplice, ogni elemento y_i della serie si suppone ottenuto per campionamento da una stessa popolazione, per esempio in un foglio di calcolo mediante la funzione `rand()` / `casuale()`:

```
yᵢ ← rand()
```

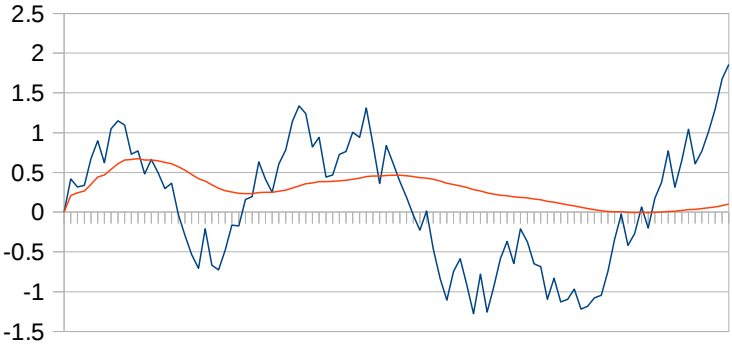
dunque nell'ipotesi che la popolazione segua una distribuzione uniforme nell'intervallo [0,1]. In questa situazione, gli elementi della serie sono generati indipendentemente l'uno dall'altro, e quindi la serie non contiene alcuna regolarità aggiuntiva rispetto a quella dovuta all'assunzione che gli elementi sono ottenuti da una stessa popolazione. Ciò si manifesta nel fatto che in questo caso la serie $\langle y_i \rangle$ ha le stesse caratteristiche statistiche di ogni altra serie ottenuta da $\langle y_i \rangle$ scambiando l'ordine dei suoi elementi. D'altra parte, serie che descrivono l'andamento nel tempo di fenomeni reali sono spesso tali che elementi successivi, y_{i+1} rispetto a y_i, non sono completamente indipendenti l'uno dall'altro, cioè appunto y_{i+1} dipende da y_i secondo una qualche funzione a meno di un termine causale. Serie di questo genere si chiamano *cammini casuali* (in inglese *random walk*). Il caso più semplice di cammino casuale è dunque $y_{i+1}=y_i+z_i$, dove z_i è appunto il termine casuale, per esempio scelto con distribuzione uniforme in un intervallo dato.
Su una serie storica si possono calcolare, naturalmente le usuali statistiche campionarie. In particolare, mediana e media rappresentano valori centrali della serie, e la deviazione standard la sua dispersione. Si possono calcolare il minimo e il massimo, corrispondenti allo zeresimo e al centesimo percentile, e ancora mediante i percentili si possono costruire gli intervalli che contengono una percentuale data dei valori della serie. Per esempio, nell'intervallo tra il quinto e il novantacinquesimo percentile è incluso il 90% degli elementi della serie (nel caso la serie sia interpretata come un campione, tale intervallo è detto *intervallo di confidenza*, e *livello di confidenza* la percentuale, perché con ciò si descrive la fiducia, la "confidenza" appunto, che gli elementi di un generico campione siano inclusi nell'intervallo stesso). Per ogni serie storica data, ognuna di queste statistiche campionarie è dunque un numero.
Il limite principale dell'uso di queste statistiche per l'analisi di serie storiche è che l'informazione portata da media, deviazione standard, percentili, ... non dipende dall'ordine degli elementi della serie. Detto altrimenti, queste statistiche sono invarianti per permutazione degli elementi del campione su cui sono calcolate. Dunque, per esempio, i campioni $\langle 10, 20, 30, 40, 50 \rangle$ e $\langle 50, 40, 30, 20, 10 \rangle$ hanno la stessa media, ma naturalmente, se interpretati come serie storiche, portano un'informazione ben diversa. Occorre dunque introdurre nuove statistiche, che tengano conto dell'ordine temporale degli elementi della serie: le statistiche progressive e le statistiche mobili.

Statistiche progressive

Una prima tecnica per l'analisi di una serie storica $\langle y_i \rangle$ consiste semplicemente nel calcolo di una o più statistiche effettuato progressivamente, cioè mentre la serie si forma nel tempo, ottenendo così una nuova serie per ogni statistica calcolata. Statistiche di questo genere si chiamano *progressive*. Consideriamo per esempio il caso della media progressiva. Mentre la serie si forma, dunque per i valori dell'indice i=1, 2, 3, ..., si può calcolare la serie $\langle y_1, (y_1+y_2)/2, (y_1+y_2+y_3)/3, ... \rangle$, tale cioè che l'$i$-esimo elemento mp_{y_i} del campione di medie progressive è:

$$\mathrm{mp}_{y_i} = \frac{\sum_{j=1}^{i} y_j}{i}$$

Il grafico che segue:



visualizza un cammino casuale $y_{i+1}=y_i+z_i$, con $y_1=0$ e z_i casuale estratto dalla distribuzione uniforme nell'intervallo [–0,5, 0,5], insieme con la sua media progressiva. Come si vede nel grafico, poiché all'aumentare del valore dell'indice la media è calcolata su un numero di elementi crescente, la sua sensibilità si riduce, cioè al crescere di i essa tende a mantenersi costante.
Con la stessa logica altre statistiche progressive, per mediana, percentili, ..., possono essere calcolate.

Statistiche mobili

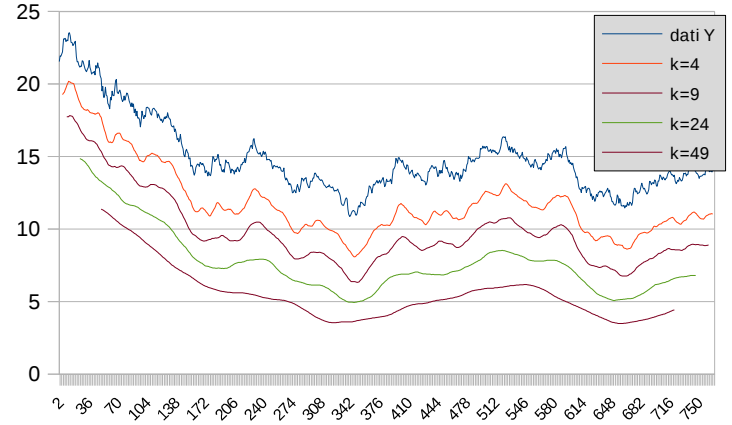
Analogamente alle statistiche progressive si possono calcolare le statistiche mobili, in cui la statistica in considerazione è calcolata su un sotto-campione di valori temporalmente contigui della serie storica di partenza. Se per esempio della serie storica $\langle y_i \rangle$ consideriamo sotto-campioni di tre elementi e vogliamo calcolarne la media mobile, otterremo la serie $\langle (y_1+y_2+y_3)/3, (y_2+y_3+y_4)/3, (y_3+y_4+y_5)/3, ... \rangle$, tale cioè che l'$i$-esimo elemento mm_{y_i} del campione di medie mobili con "finestra di osservazione" di ampiezza 3 è:

$$\mathrm{mm}_{y_i} = \frac{\sum_{j=i-2}^{i} y_j}{3}$$

dove dunque l'estremo destro della finestra di osservazione è proprio l'istante i-esimo, considerato come il tempo presente, che si sposta progressivamente nel corso della costruzione della serie storica.
Un'interessante applicazione delle medie mobili è la seguente. Certe serie storiche si suppongono costituite da un "segnale" a cui risulta sovrapposto del "rumore", variabile in modo casuale e che si vorrebbe eliminare, dunque "separando il segnale dal rumore", con un'operazione di *smoothing* ("lisciatura"). Data la serie storica $\langle y_i \rangle$, una tecnica semplice a questo scopo consiste nel generare una nuova serie, in cui ogni y_i è sostituito dalla media mobile mm_{y_i}, calcolata questa volta come:

$$\mathrm{mm}_{y_i} = \frac{\sum_{j=i-k}^{i+k} y_j}{2k+1}$$

dunque calcolando le medie via via su finestre di osservazione costituite da un numero costante $2k+1$, con $k \geq 1$, di valori e centrate sul valore i-esimo: tale nuovo campione si chiama di *medie mobili centrate*. Come si vede nella figura, ampliando la dimensione della finestra, si ottengono successioni "sempre più lisce" (ATTENZIONE: le diverse serie di medie mobili sono traslate lungo l'asse y solo per chiarezza di visualizzazione):

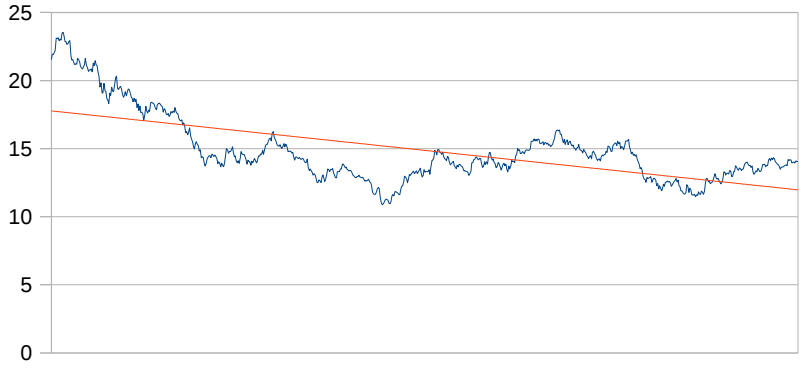


Analisi di trend: regressione lineare

In una serie storica si può riconoscere in particolare la presenza di un trend e di periodicità (chiamate a volte "stagionalità"). Nel caso più semplice, il trend si può intendere come una retta $w_i=\beta_1 t_i+\beta_2$ su cui si ipotizza si dispongano i valori y_i, a meno di un termine additivo ε_i, dunque:

$$y_i=(\beta_1 t_i+\beta_2)+\varepsilon_i$$

come nell'esempio:



Secondo questa ipotesi, detta di *regressione lineare*, dato il campione $\langle y_i \rangle$, si tratta dunque di trovare la miglior retta per descrivere il trend della serie storica, e quindi i migliori valori β_1 e β_2 per i parametri della retta, secondo un qualche criterio di ottimalità. Chiamati *residui* i termini ε_i tali che:

$$\varepsilon_i=y_i-w_i=y_i-(\beta_1 t_i+\beta_2)$$

il criterio abituale, chiamato "dei minimi quadrati" (in inglese: *least squares*), calcola i parametri β_1 e β_2 in modo da minimizzare il valore:

$$\sum_{i=1}^{n} \varepsilon_i^2$$

cioè appunto la somma quadratica dei residui (si noti che, per costruzione, la somma dei residui è nulla:

$$\sum_{i=1}^{n} \varepsilon_i=0$$

e quindi lo è in particolare la loro media). La soluzione del problema:

$$\min_{\beta_1,\beta_2} \sum_{i=1}^{n} (y_i-(\beta_1 t_i+\beta_2))^2$$

che ha dunque la forma:

$$\min_{\beta_1,\beta_2} f(\beta_1,\beta_2)$$

richiede competenze analitiche più sofisticate di quelle richieste qui.

> D'altra parte, i fogli di calcolo mettono a disposizione le funzioni `slope(y,x)` e `intercept(y,x)` per ottenere i valori dei parametri β_1 e β_2 a partire da un campione $\langle\langle t_i,y_i \rangle\rangle$, e quindi con ciò possiamo considerare risolto il nostro problema: dati i valori osservati y_i, con un criterio di minimizzazione abbiamo identificato dei "valori di modello" $w_i=\beta_1 t_i+\beta_2$.

Possiamo chiederci quanto sia buono il modello fornito dalla retta di regressione rispetto ai valori osservati. Intuitivamente, minori sono i residui migliore è il modello. Su questa base, diamo due criteri che, si può dimostrare, producono esattamente lo stesso risultato, il *coefficiente di determinazione* R^2 che ha valore tanto maggiore quanto migliore è il modello.

Il primo criterio. Consideriamo la varianza della serie dei residui, s_ε^2: quanto minore è, tanto migliore è il modello. Definiamo allora:

$$R^2=1-\frac{s_\varepsilon^2}{s_y^2}$$

in cui s_ε^2 è normalizzata rispetto alla varianza della serie di partenza, s_y^2.

Il secondo criterio. Consideriamo la correlazione tra la serie dei valori di modello, $\langle w_i \rangle$, e la serie di partenza $\langle y_i \rangle$, $R_{y,w}$: quanto maggiore è, tanto migliore è il modello. Definiamo allora:

$$R^2=R_{y,w}^2$$

Da questa espressione, e ricordando che il coefficiente di correlazione campionaria ha valori nell'intervallo $[-1,1]$, si conclude che il coefficiente di determinazione R^2 ha valori tra 0 e 1.

> I fogli di calcolo mettono a disposizione la funzione `rsq(x,y)` per ottenere il valore del coefficiente di determinazione R^2.

Una volta che sia stata calcolata, la retta di regressione può essere usata in particolare per:

- *interpolazione*: dato un valore t non compreso tra quelli della serie storica ma interno all'intervallo $[t_1, t_n]$, si può calcolare il valore $w = w(t)$ che giace sulla retta di regressione, $w=\beta_1 t+\beta_2$: si tratta del più semplice esempio di un valore calcolato per interpolazione;
- *estrapolazione*: dato un valore t esterno all'intervallo $[t_1, t_n]$, e in particolare successivo a t_n, si può calcolare il valore $w = w(t)$ che giace sulla retta di regressione, $w=\beta_1 t+\beta_2$: si tratta del più semplice esempio di un valore calcolato per estrapolazione; nel caso in cui t_n sia l'istante presente, l'estrapolazione consente di effettuare una previsione su un valore futuro della serie storica, a partire dal suo trend, e per questo motivo i valori di modello w_i sono anche chiamati predittori.

Una nota sul metodo dei minimi quadrati

Il metodo dei minimi quadrati, che abbiamo impiegato per determinare i parametri della retta di regressione, è uno strumento assai generale, e in linea di principio non è particolarmente complesso da usare. Per vederlo in azione in un caso sufficientemente semplice, dimostriamo che la media aritmetica m_y di un generico campione $\langle y_i \rangle$ di n elementi è il valore z che minimizza la somma $\sum_i (y_i - z)^2$.

Cerchiamo dunque $\min_z \sum_i (y_i - z)^2 = \min_z (\sum_i y_i^2 - \sum_i 2\, y_i z + \sum_i z^2)$. Troviamo il minimo della funzione calcolando gli zeri della sua derivata prima. Dunque $\frac{d}{dy}(\sum_i y_i^2 - \sum_i 2\, y_i z + \sum_i z^2) = 0$, da cui $2\, n z = \sum_i 2\, y_i$, e quindi $z = \frac{\sum_i 2\, y_i}{2\, n} = \frac{\sum_i y_i}{n}$.

Regressioni non lineari

In certe situazioni la regressione lineare non appare il modello migliore per interpretare i dati disponibili, come potrebbe mettere in evidenza un basso valore del coefficiente di determinazione R^2. Sulla base di una logica analoga a quella impiegata finora, si può allora adottare un modello per la regressione non lineare della serie storica.

Consideriamo il caso della *regressione logaritmica*: invece di una retta $w=\beta_1 t+\beta_2$, cerchiamo una funzione $w=\beta_1 \ln(t)+\beta_2$, con parametri β_1 e β_2 opportuni, in grado di fornire una buona regressione della serie $\langle y_i \rangle$ di partenza. Il problema si riconduce immediatamente a quello di regressione lineare con il cambio di variabile $\ln(t) \to t'$: nuovamente attraverso il metodo dei minimi quadrati, si tratta dunque di trovare i parametri β_1 e β_2 che minimizzano la somma dei residui:

$$\varepsilon_i = y_i - w_i = y_i - (\beta_1 \ln(t_i) + \beta_2) = y_i - (\beta_1 t'_i + \beta_2)$$

elevati al quadrato.

> Come abbiamo visto, nei fogli di calcolo sono disponibili le funzioni `slope(y,x)` e `intercept(y,x)` per risolvere il problema di ottenere i valori dei parametri β_1 e β_2. In questo caso tali funzioni devono essere dunque applicate al campione $\langle \ln(t_i), y_i \rangle$.
> Il coefficiente di determinazione R^2 si può poi calcolare, come nel caso precedente, come quadrato del coefficiente di correlazione campionaria di $\langle y_i \rangle$ e $\langle w_i \rangle$, oppure anche – ricordando che esso è invariante per trasformazioni lineari e che $w_i = \beta_1 \ln(t_i) + \beta_2$ – di $\langle y_i \rangle$ e $\langle \ln(t_i) \rangle$.

Solo un poco più complesso è il caso della *regressione esponenziale*, in cui il regressore è una funzione del tipo $w=\beta_2 \exp(\beta_1 t)$, con parametri β_1 e β_2 opportuni. Per ricondurre anche questo a un problema di regressione lineare, calcoliamo il logaritmo di entrambi i termini dell'equazione, $\ln(w) = \ln(\beta_2 \exp(\beta_1 t))$, e perciò $\ln(w) = \ln(\beta_2) + \beta_1 t$. Effettuando i due cambi di variabile $\ln(w) \to w'$ e $\ln(\beta_2) \to \beta_2'$, si ottiene la funzione lineare $w' = \beta_1 t + \beta_2'$. Il metodo dei minimi quadrati sarà dunque da applicare per minimizzare:

$$\varepsilon_i^2 = (\ln(y_i) - w_i')^2$$

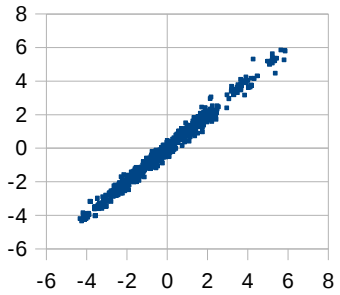
e, una volta ottenuti β_1 e β_2', calcolando quindi $\beta_2 = \exp(\beta_2')$.

In questo caso le funzioni `slope(y,x)` e `intercept(y,x)` devono essere dunque applicate al campione $\langle\langle t_i, \ln(y_i)\rangle\rangle$, e analogamente il coefficiente di determinazione deve essere calcolato come quadrato del coefficiente di correlazione campionaria di $\langle \ln(y_i)\rangle$ e $\langle \ln(w_i)\rangle$ oppure, data la dipendenza lineare di $\ln(w)$ da t, di $\langle \ln(y_i)\rangle$ e $\langle t_i\rangle$.

Analisi dei residui: autocorrelazione

Considerando la retta di regressione come la componente di trend della serie storica, è utile analizzare se la serie storica dei residui $\langle \varepsilon_i \rangle$ sia completamente casuale o se, al contrario, mantenga delle regolarità interne. L'idea di base è semplice: supponendo che un generico valore y_i sia per esempio al di sopra del trend, cioè $y_i > (\beta_1 t_i + \beta_2)$, e quindi che il corrispondente residuo ε_i sia positivo, se la successione fosse casuale il residuo successivo, ε_{i+1}, dovrebbe essere indifferentemente positivo o negativo; se, al contrario, un residuo positivo è frequentemente seguito da un residuo positivo, e un residuo negativo è frequentemente seguito da un residuo negativo, la successione dei residui non è completamente casuale, e dunque, in particolare, potrebbe essere almeno in parte prevedibile.

Per acquisire questa informazione si può dunque confrontare ε_1 con ε_2, ε_2 con ε_3, e così via, cioè si può confrontare la serie storica dei residui $\langle \varepsilon_i \rangle$ con una sua copia ritardata di un elemento, che possiamo indicare per convenzione con $\langle \varepsilon_{i+1} \rangle$. Per questo confronto si fa uso del coefficiente di correlazione campionaria, $r(\langle\langle \varepsilon_i, \varepsilon_{i+1}\rangle\rangle)$, chiamato in questo caso *coefficiente di autocorrelazione*, anch'esso naturalmente a valori nell'intervallo $[-1,1]$, potendo inoltre rappresentare il campione $\langle\langle \varepsilon_i, \varepsilon_{i+1}\rangle\rangle$ su un diagramma di dispersione:



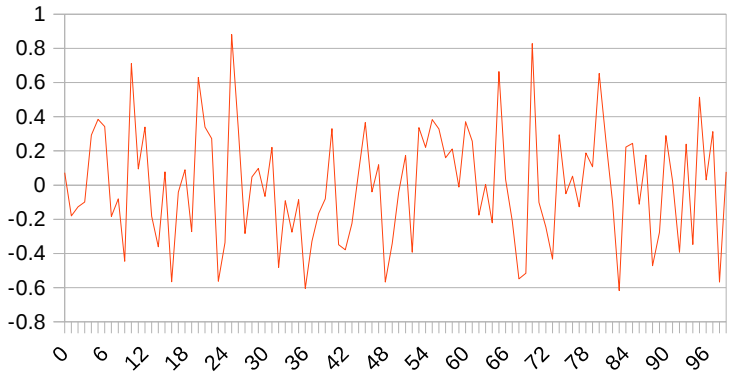
(nella serie dell'esempio precedente: autocorrelazione: 0,993) (perché l'autocorrelazione è calcolata sulla serie dei residui invece che sulla serie storica di partenza? la risposta dovrebbe essere chiara: se la serie storica ha un trend, e quindi non è tempo-stazionaria, una componente di autocorrelazione dipende dalla presenza del trend).

Analisi dei residui: autocorrelogramma

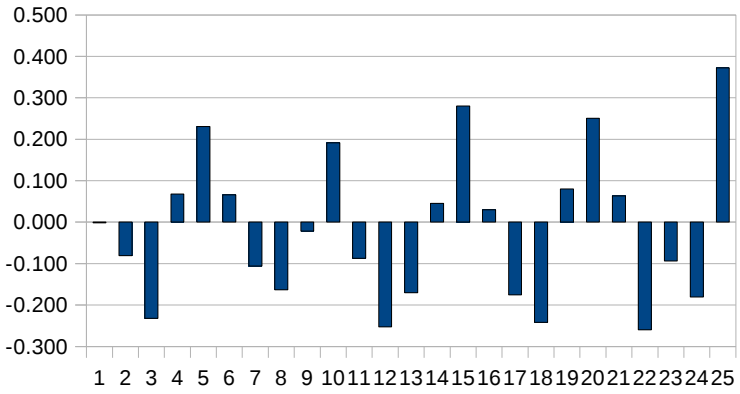
L'autocorrelazione dei residui può essere usata anche per cercare delle eventuali periodicità nel campione di partenza. A questo scopo la successione dei residui $\langle \varepsilon_i \rangle$ può essere confrontata successivamente con sue copie ritardate di uno, due, tre, ... elementi, consentendo perciò di calcolare i coefficienti di correlazione campionaria, $r(\langle\langle \varepsilon_i, \varepsilon_{i+1}\rangle\rangle)$, $r(\langle\langle \varepsilon_i, \varepsilon_{i+2}\rangle\rangle)$, $r(\langle\langle \varepsilon_i, \varepsilon_{i+3}\rangle\rangle)$, ...

Il fatto che $r(\langle\langle \varepsilon_i, \varepsilon_{i+k}\rangle\rangle)$ sia sufficientemente prossimo a 1 per un certo valore k, e quindi che se $\langle \varepsilon_i \rangle$ è positivo anche $\langle \varepsilon_{i+k} \rangle$ sistematicamente lo è, mette in evidenza la presenza nel campione di una componente di periodicità t_k. La successione dei valori $r(\langle\langle \varepsilon_i, \varepsilon_{i+k}\rangle\rangle)$ può essere rappresentata graficamente in funzione della successione $\langle t_k \rangle$, per esempio mediante un istogramma, ottenendo un diagramma chiamato *autocorrelogramma*.

Considerando per esempio la serie storica:



il suo autocorrelogramma è:



e mette in evidenza (cosa che non era così chiara nel grafico della serie) una periodicità di periodo k=5 (e quindi naturalmente anche dei suoi multipli, k=10, k=15, ...).

Un esempio: controllo statistico di processo

Molti processi di produzione industriale sono ripetitivi, nel senso che a intervalli di tempo (generalmente) regolari producono un prodotto la cui qualità è condizionata dalla conformità del prodotto a specifiche date.
Un processo ideale ha dunque una perfetta ripetibilità, cioè i suoi prodotti sono tutti uguali (ed evidentemente tutti conformi alle specifiche). Nella realtà, ogni processo è soggetto a una certa variabilità: ogni grandezza che viene misurata sui prodotti (tutti o solo alcuni, secondo una qualche logica campionaria) produce dunque una successione di misure può essere allora interpretata come una serie storica, $x = x(t)$, con t che varia sull'insieme (tipicamente discreto) degli istanti in cui la grandezza x viene misurata (si noti che l'enfasi sul processo, invece che sui prodotti, fa sì che il riferimento ai prodotti stessi venga tralasciato, e quindi si scriva appunto $x(t)$ come forma abbreviata di $x(p(t))$ (la grandezza x misurata sul prodotto p al tempo t). Per l'ipotesi di tempo discreto, si può infine scrivere x_i al posto di $x(t)$, dunque con l'indice i che varia nel tempo.
Il controllo statistico di processo opera dunque su serie storiche tali che:

- quando le cause di tale variabilità sono solo casuali ("cause naturali", ingl. *common cause*), si può ipotizzare che nella gran parte dei casi i valori delle grandezze siano nei limiti di conformità; in questo caso si dice che il processo è *"sotto controllo"*;
- quando invece la variabilità è causata da fattori non casuali (potremmo dire: sistematici; ingl. *special cause*) il processo è considerato *"fuori controllo"*.

In questo contesto, le carte di controllo (ce ne sono di diversi tipi) sono state introdotte come strumento di gestionale, sulla base del principio che la qualità di processo è interpretabile come bassa variabilità. La carta di controllo usata più frequentemente è un diagramma cartesiano bidimensionale tale che:

- in ascissa si mette il tempo di rilevazione;
- in ordinata si mette la grandezza in esame.

Per facilità di uso, in ordinata vengono anche tracciate delle linee di riferimento che, nell'ipotesi di grandezza con specifiche bilaterali, sono:

- la linea centrale (CL) per la media del processo $\bar{x}$ (che si suppone allineato con il valore nominale, dunque nell'ipotesi di assenza di fattori sistematici);

- le linee di controllo (UCL e LCL), tipicamente a $\pm 3\sigma_{\bar{x}}$ (quindi date la deviazione standard σ_X del processo e la dimensione n dei campioni considerati, $\sigma_{\bar{x}}=\sigma_X/\sqrt{n}$, naturalmente nell'ipotesi di indipendenza statistica tra elementi dei campioni);
- le eventuali linee di "sorveglianza" (UWL e LWL, *upper and lower warning limits*), per esempio stabilite a $\pm\sigma_{\bar{x}}$ e $\pm 2\sigma_{\bar{x}}$, per anticipare le eventuali azioni correttive.

Se un processo è sotto controllo tutti i punti della carta di controllo sono entro i limiti di controllo con alta probabilità (almeno $1-1/k^2$ per la disuguaglianza di Chebyshev, e quindi, dato $k=3$, almeno $8/9=0.89$; nell'ipotesi che il processo segua una distribuzione gaussiana, 0.997).
Per cercare di individuare rapidamente delle condizioni di processo plausibilmente fuori controllo, si adottano delle regole empiriche, per esempio:

- regola 1: un punto fuori dai limiti $\pm 3\sigma_{\bar{x}}$;
- regola 2: 2 su 3 punti consecutivi fuori dal limite $\pm 2\sigma_{\bar{x}}$ nello stesso lato rispetto alla CL;
- regola 3: 4 su 5 punti consecutivi fuori dal limite $\pm\sigma_{\bar{x}}$ nello stesso lato rispetto alla CL;
- regola 4: 9 punti consecutivi nello stesso lato rispetto alla CL.

La logica delle carte di controllo è fondata sul tentativo di individuare delle regolarità nella successione x_i, ipotizzate come causate da fattori non casuali e quindi indice di processo fuori controllo. Da un punto di vista statistico, tali regolarità si manifestano in termini di trend non nullo o di autocorrelazione (significativamente) non nulla della successione x_i.
Evidentemente, tutto ciò non è comunque nient'altro che un caso particolare del concetto generale di qualità come conformità a specifiche, stabilita confrontando "quello che dovrebbe essere" (sulla carta) con "quello che è" (nella pratica).

Correlazione e trend

Il metodo dei minimi quadrati per la ricerca dei parametri della retta di regressione è in effetti applicabile non solo a serie storiche ma anche, e più in generale, a campioni bivariati $\langle\langle x_i, y_i\rangle\rangle$. In tal caso è utile studiare la relazione tra il coefficiente di correlazione campionaria $r_{x,y}$ e la pendenza $\beta_{y,x}$ della retta che definisce il trend del campione $\langle\langle x_i, y_i\rangle\rangle$. Vale in generale l'interessante risultato che:

$$\beta_{y,x}=r_{y,x}\frac{s_y}{s_x}$$

dove, come in precedenza, s_x e s_y sono le deviazioni standard dei campioni $\langle x_i\rangle$ e $\langle y_i\rangle$ rispettivamente.
Un secondo interessante risultato si ottiene attraverso la trasformazione cosiddetta di standardizzazione dei due campioni tale da costruire da $\langle x_i\rangle$ un nuovo campione:

$$\langle x'_i\rangle=\langle\frac{x_i-m_x}{s_x}\rangle$$

a analogamente per $\langle y_i\rangle$. E' facile vedere che un campione in questo senso standardizzato ha media $m_{x'}=0$ e deviazione standard $s_{x'}=1$. A seguito della standardizzazione di $\langle x_i\rangle$ e $\langle y_i\rangle$, il rapporto $s_{y'}/s_{x'}$ è dunque pari a 1, e perciò:

$$\beta_{y',x'}=r_{y',x'}$$

Nel caso le due variabili di un campione bivariato siano standardizzate, la pendenza della retta di trend coincide con il coefficiente di correlazione.