Lezione 5

Trasformazione delle variabili

Stefania Capecchi

Università degli Studi di Napoli Federico II

stefania.capecchi@unina.it

Sintesi della lezione

1. ***Motivazioni per la trasformazione delle variabili***
2. ***Tipologia delle trasformazioni***
3. ***Trasformazioni matematiche***
4. ***Trasformazioni statistiche***
5. ***Applicazioni***

Perché trasformare una variabile

- Molte delle distribuzioni che si applicano sono utili (e talora necessarie) per rendere ottimali alcune analisi inferenziali di cui si discuterà più avanti.

- Quando i dati sono asimmetrici e/o la distribuzione presenta un forte appuntimento, si possono applicare ***trasformazioni normalizzanti*** allo scopo di rendere la distribuzione delle modalità più simile a quella di una distribuzione Gaussiana (di cui si parlerà più oltre).

- Altre trasformazioni sono introdotte per rendere più visibili talune caratteristiche dei dati ovvero, nella fase esplorativa, per lavorare su dati maggiormente omogenei.

- Sono gli ***obiettivi dell'analisi*** a giustificare ed a richiedere una particolare trasformazione. Per questo, *non esistono trasformazioni sempre ottimali*.

- Quasi sempre, alla fine delle analisi statistiche sulle variabili trasformate è necessarieo ritornare ai dati originari per presentare e commentare i risultati tramite misure comprensibili ed utili.

Classificazione delle trasformazioni

- Le principali trasformazioni sono:
  - di tipo *matematico*, perché introdotte per sfruttare le proprietà analitiche di qualche funzione matematica
  - di tipo *statistico*, perché introdotte per ottimizzare proprietà statistiche dei dati

- La distinzione non va intesa in senso rigido perché ogni trasformazione statistica è, in pratica, una trasformazione matematica e tutte le trasformazioni matematiche vengono introdotte per finalità anche statistiche.

Trasformazioni di tipo matematico

➤ Tra le principali trasformazioni di tipo matematico, segnaliamo:

- ***lineare***
- ***potenza***, ***radice***, etc.
- ***logaritmica***
- ***logit***
- ***arc-seno***
- . . . . . .

Trasformazione lineare

➤ Se X è una variabile quantitativa ed a e b sono due costanti reali note, allora una ***trasformazione lineare*** di X è definita da:

$$X^* = aX + b\,.$$

➤ Una trasformazione lineare modifica la posizione di X traslando tutte le modalità di b e moltiplicando/dividendo ciascuna di esse per a.

➤ É utile per modifiche dell'unità di misura, se le modalità sono numeri troppo grandi/piccoli, per confrontare valori espressi in monete differenti, etc.

➤ Una trasformazione lineare modifica la posizione e la variabilità di X ma non la forma della distribuzione.

Trasformazioni potenza

- Una trasformazione ***potenza/radice*** della variabile quantitativa X che assume modalità non-negative, per un numero reale c, è definita da:

$$X^* = X^c \, .$$

- Se $c > 0$ trattasi di una *trasformazione potenza* e la variabile X^* presenta modalità più grandi per modalità $x_i > 1$ ovvero più piccole per modalità $x_i < 1$.

- Per potenze negative: $X^{-c} = 1/X^c$ per cui si ha una *trasformazione inversa*.

- Se c è una frazione propria trattasi di una *trasformazione radice*.

- Le trasformazioni potenza modificano posizione, variabilità e forma di una distribuzione statistica.

- Una generalizzazione della trasformazione potenza è la *trasformazione Box-Cox*, che include anche la trasformazione logaritmica come caso particolare.

Trasformazioni radice

- Le trasformazioni più diffuse sono:

$$X^{\frac{1}{2}} = \sqrt{X}; \quad X^{\frac{1}{3}} = \sqrt[3]{X}; \quad X^{\frac{-1}{2}} = \frac{1}{\sqrt{X}}; \ldots$$

- La trasformazione *radice quadrata* è spesso utilizzata per studiare variabili che esprimono conteggi di eventi: incidenti, episodi traumatici, visite, numero di auto in fila ad un casello autostradale, etc.

- La trasformazione *radice cubica* è utilizzata nello studio di fenomeni idrologici, come per esempio la piovosità.

Trasformazione logaritmica...1

➤ La ***trasformazione logaritmica*** della variabile quantitativa e non-negativa X si ottiene mediante:

$$X^* = \log(X)$$

dove $\log(.)$ è il logaritmo in qualsiasi base.

➤ Più frequentemente, si usa la base e per i logaritmi neperiani (dotati di maggiori proprietà analitiche) oppure la base 10 per i logaritmi decimali (di più immediata interpretazione).

➤ Qualunque sia la base, l'effetto della trasformazione logaritmica è lo stesso, a meno di un coefficiente di proporzionalità.

Trasformazione logaritmica...2

- L'effetto della trasformazione logaritmica è quello di ridurre la distanza tra la modalità minima e quella massima, il che rende più simmetrica la distribuzione della variabile avvicinando i valori estremi a quelli centrali.
- Per esempio, le modalità $\{10, 100, 1000, 10000\}$ sono tali che il massimo è 1000 volte più grande del minimo.
- Se si opera la trasformazione logaritmica (in base 10, per semplicità) le modalità diventano: $\{1, 2, 3, 4\}$. Ora, il massimo è solo 4 volte più grande del minimo.
- Inoltre, il campo di variazione di X era di $10000 - 10 = 9990$ mentre il campo di variazione di $\log(X)$ si è ridotto a $4 - 1 = 3$.

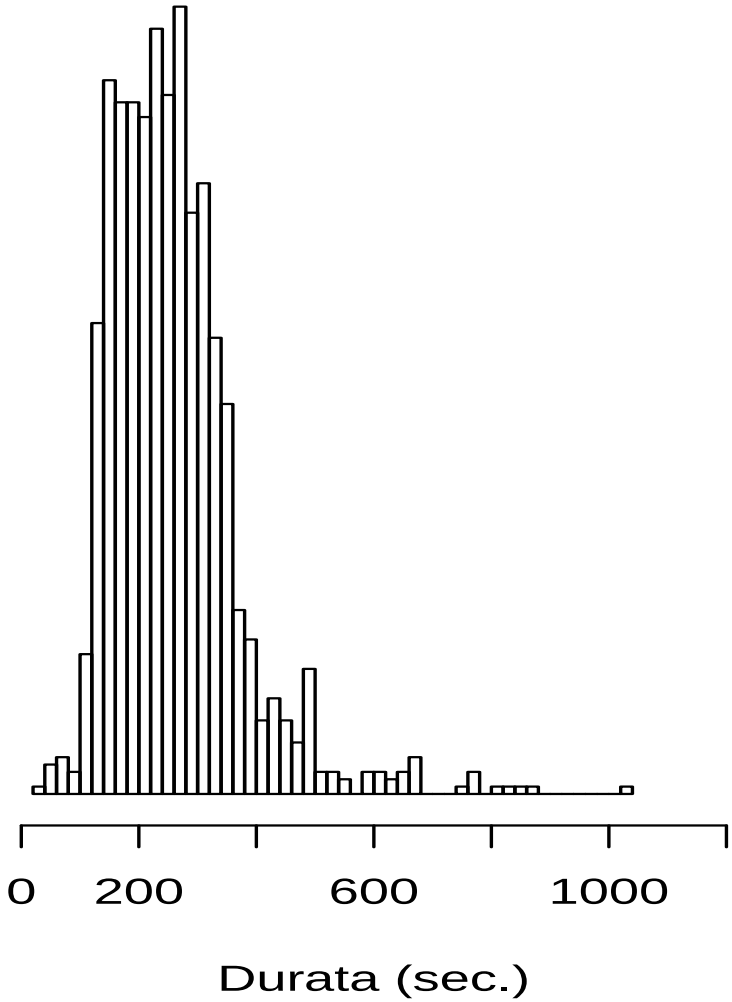
Trasformazione logaritmica...3

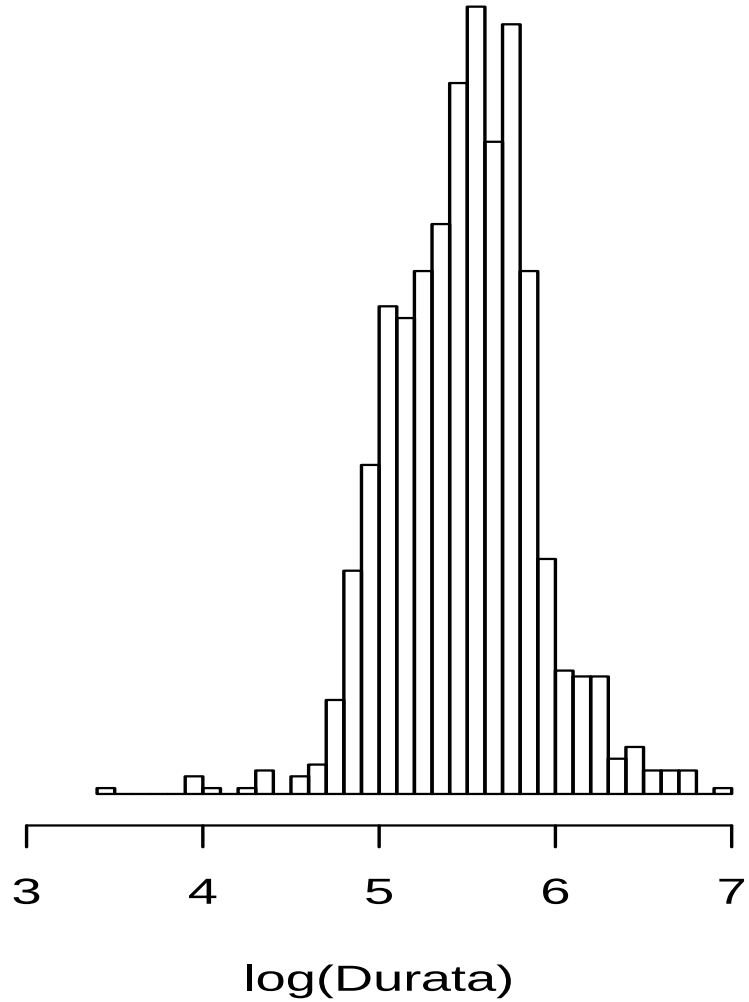
- La trasformazione logaritmica è molto utilizzata nello studio della distribuzione del reddito, dei consumi, di molte variabili economiche, della durata delle conversazioni telefoniche, etc. perché riconduce distribuzioni asimmetriche positive ad una forma più simmetrica e con code ridotte e più bilanciate.

- In presenza di pochi valori negativi, la trasformazione logaritmica si può ancora applicare considerando la variabile $\log(X + d)$ dove d è un numero positivo maggiore in valore assoluto del più piccolo valore negativo presente nei dati.

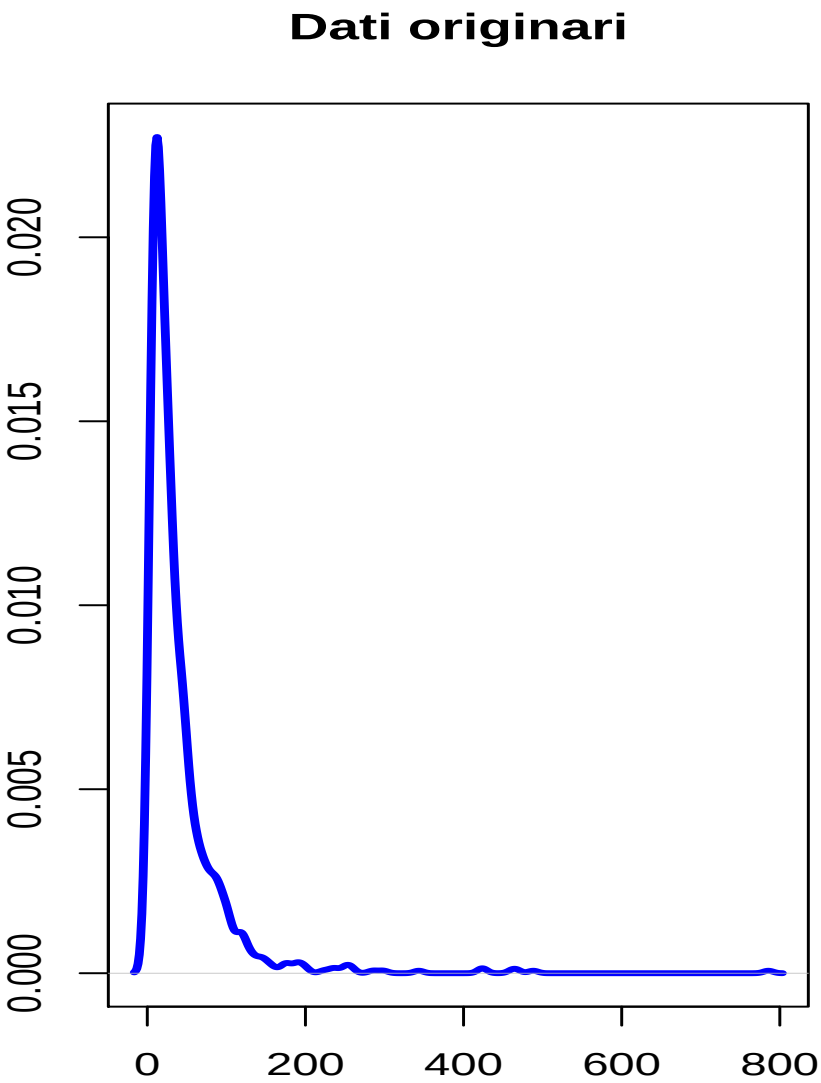
- Per esempio, ad una variabile X con modalità $\{-10, -7, 10, 11, 18, 24\}$ non si può applicare la trasformazione logaritmica. Tuttavia, essendo -10 il valore negativo più piccolo e scegliendo, per esempio, $d = 11$ è ben definita la trasformazione $\log(X + 11)$ perché la variabile $X + 11$ assume le modalità $\{1, 4, 21, 22, 29, 35\}$ che sono tutte non-negative.

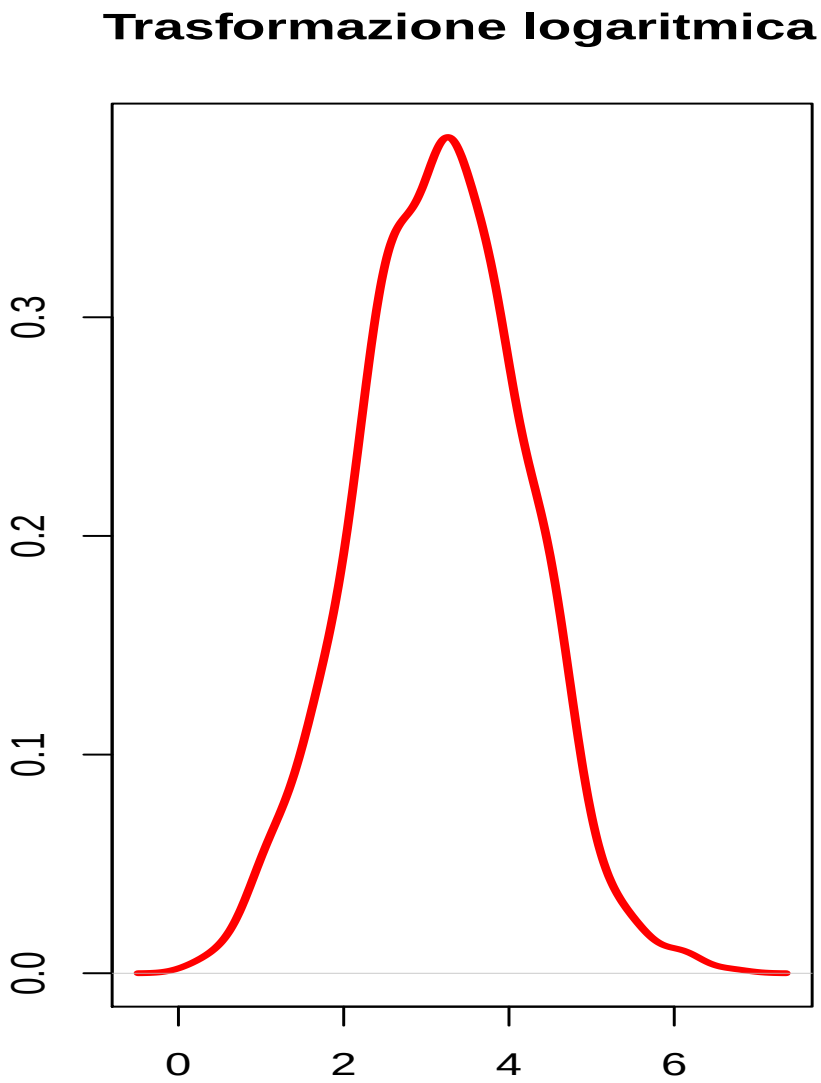
Dati originari e trasformazione logaritmica...1





Dati originari e trasformazione logaritmica...2





Effetto della trasformazione logaritmica

➤ La tabella presenta alcuni indici statistici per la distribuzione della variabile $X =$"Durata in secondi" di $n = 1192$ brani musicali. La variabile $\log(X)$ indica il logaritmo neperiano (naturale) di x.

Indici statistici	Simboli	*Variabile* X	*Variabile* $\log(X)$
Minimo	$x_{(1)}$	30.000	3.401
1° Quartile	Q_1	182.000	5.204
Mediana	Me	244.000	5.497
Media	μ	259.336	5.480
3° Quartile	Q_3	308.000	5.730
Massimo	$x_{(n)}$	1022.000	6.930
Scarto quadratico medio	σ	109.668	0.395
Coefficiente di variazione	CV	0.429	0.072
Campo di variazione	$x_{(n)} - x_{(1)}$	992.000	3.528
InterQuartile Range	IQR	126.000	0.526
InterDecile Range	IDR	225.000	0.933
Indice asimmetria	$\mathcal{A}$	0.140	-0.044
Coefficiente di asimmetria	γ	1.824	-0.094
Coefficiente di curtosi	β	9.336	4.298

Trasformazione logit

➤ La ***trasformazione logit*** (detta anche *logistica*) si applica a modalità $x \in (0, 1)$ ed è definita da:

$$X^* = \mathit{logit}(X) = \log\left(\frac{X}{1-X}\right) .$$

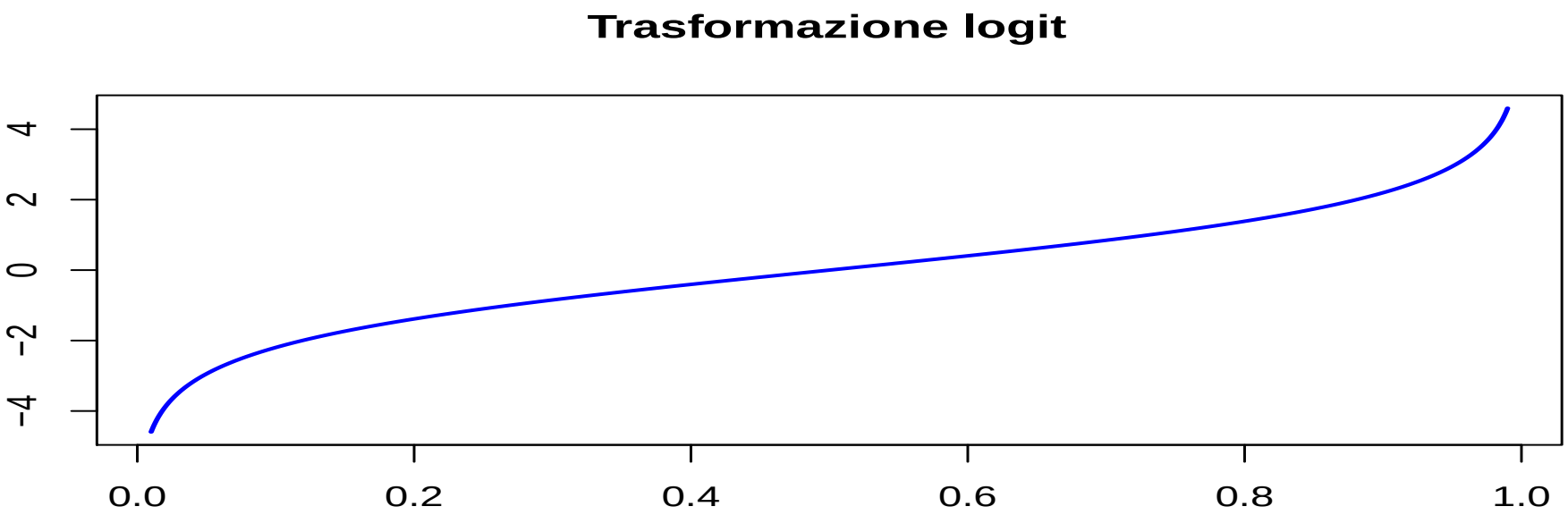
➤ Dopo agevoli passaggi, si deriva la *trasformazione logit inversa*:

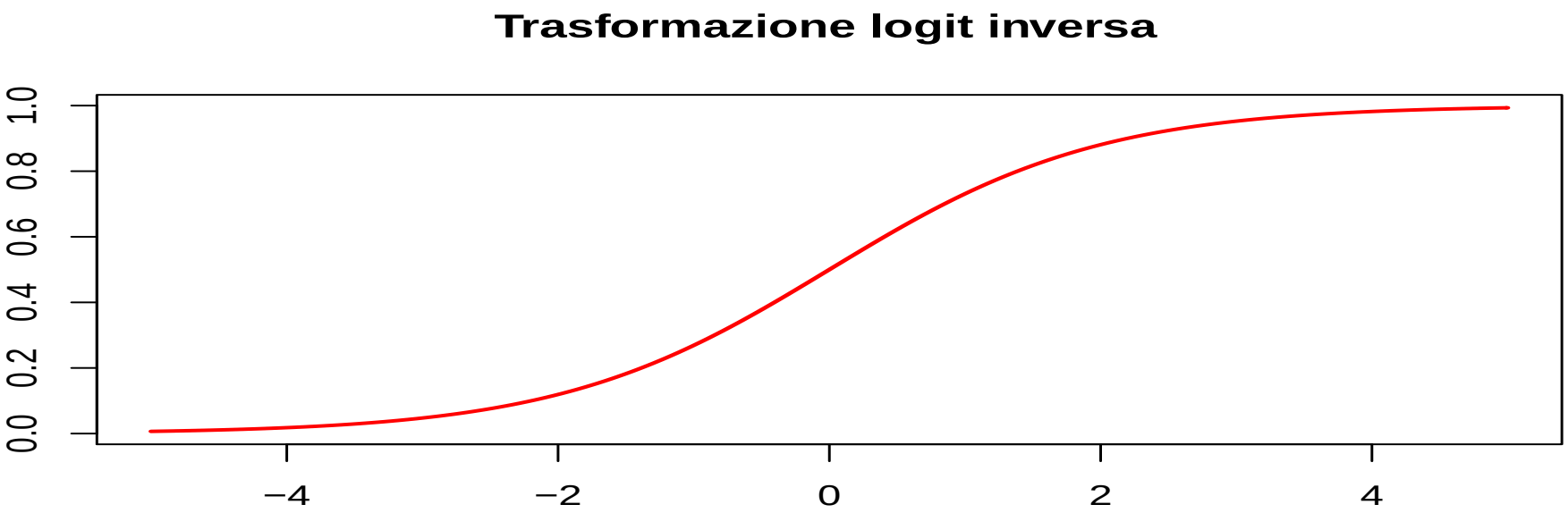
$$X = \mathit{logit}^{-1}(X^*) = \frac{1}{1+e^{-X^*}} .$$

➤ La variabile trasformata è nulla se $X = 1/2$, è positiva (negativa) per $X > 1/2$ (per $X < 1/2$).

➤ Per valori di $X \in (0.3, 0.7)$ l'effetto della trasformazione logistica è quasi lineare, per cui la modifica sostanziale della distribuzione avviene nelle due code.

Trasformazione logit e logit inversa



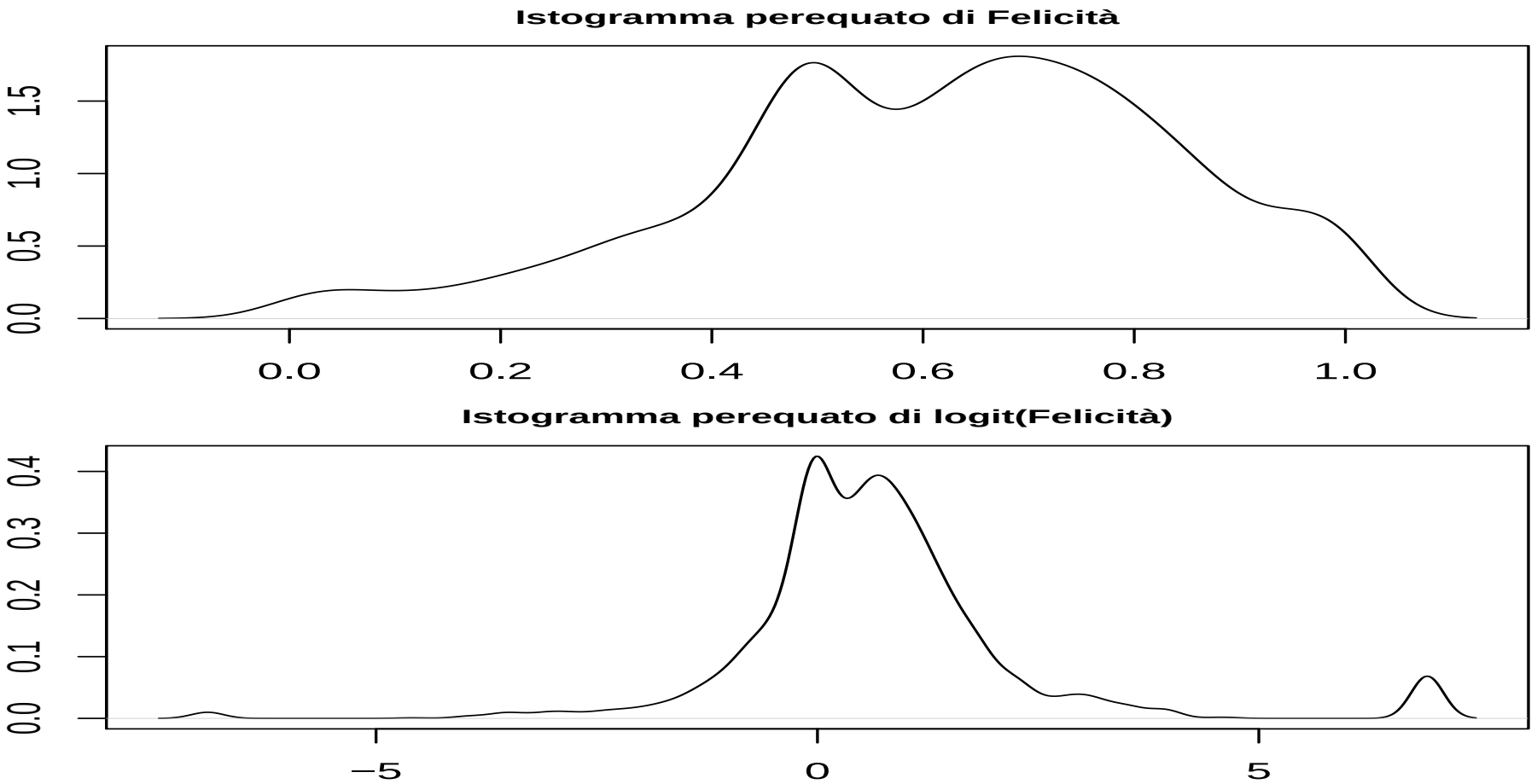


Utilizzo della trasformazione logit

- La trasformazione logistica trasforma valori compresi fra 0 e 1 in numeri compresi su tutto l'asse reale.
- Per questo, si rileva utile quando si dispone di probabilità, frazioni proprie, proporzioni, quote che –per loro natura– non possono essere utilizzate per le tecniche statistiche che prevedono valori su tutto l'asse reale (come quando si assume l'ipotesi della Normalità per X).
- Se nei dati, sono presenti modalità pari a 0 e/o 1 la trasformazione logit produce valori pari a $-\infty$ e $+\infty$, rispettivamente. Per ovviare a questi casi si possono sostituire i valori estremi con valori assai vicini (per esempio, 0 con 0.0001 e 1 con 0.9999).
- Se valori nulli o unitari sono molto diffusi si suggerisce allora la *trasformazione arc-seno*.

Trasformazione logit per la variabile "Happiness"

➤ La variabile "Happiness" è stata costruita normalizzando sull'intervallo $[0, 1]$ le dichiarazioni di un campione di rispondenti. I grafici mostrano l'effetto della trasformazione *logit* su tale variabile.



Trasformazione arc-seno

➤ La *trasformazione arc-seno* (che dovrebbe essere chiamata *arc-seno della radice quadrata*) della variabile X compresa fra 0 e 1 è definita da:

$$X^* = \textit{arc-sen}\left(\sqrt{X}\right).$$

➤ Tale trasformazione genera valori compresi fra 0 e $\pi/2$. Anche per la trasformazione arc-seno, l'effetto della trasformazione per valori di $X \in (0.3, 0.7)$ è quasi lineare, per cui la sua utilità riguarda ciò che avviene nelle due code della distribuzione dei dati.

➤ Introdotta per dati generati da variabili casuali Binomiali, è molto utile nelle situazioni (come nella distribuzione di specie, in Ecologia, per esempio) in cui sono frequenti proporzioni pari a 0 (totale assenza del fenomeno) oppure pari a 1 (presenza costante del fenomeno) perché, a differenza della trasformazione logit, la trasformazione arc-seno genera valori finiti anche agli estremi.

Confronto fra trasformazioni

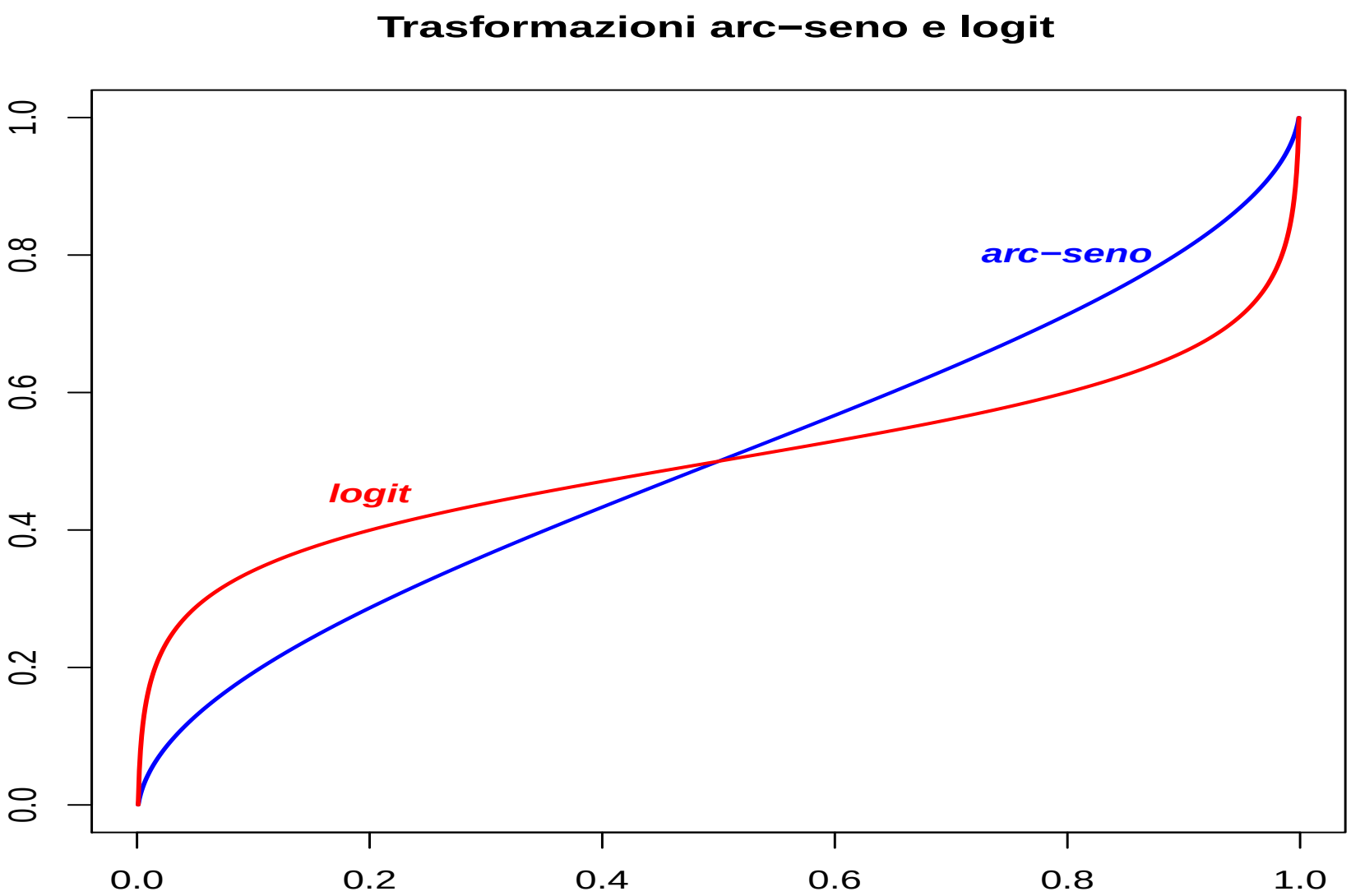
- La trasformazione arc-seno è stata introdotta poiché, soprattutto nell'analisi della varianza (ANOVA), si cerca di lavorare con osservazioni omogenee in varianza. Ora, dati generati da variabili casuali Binomiali con parametri n e θ sono caratterizzati da un valore medio pari a $n\theta$ ed una varianza pari a $n\theta(1-\theta)$. Quindi, la varianza cambia con il valore medio essendo $\sigma^2 = \mu\,(1-\theta)$.

- Per contro, è stato dimostrato che la trasformazione arc-seno genera dati con una varianza approssimativamente pari a $\frac{1}{4n}$, cioè indipendente dai parametri.

- Recenti lavori hanno criticato l'uso della trasformazione arc-seno a favore di quella logit la quale consente migliori interpretazioni e fornisce un effetto più accentuato negli estremi della distribuzione.

- Dopo aver normalizzato entrambe le trasformazioni per renderle comparabili, la Figura seguente mostra il loro differente effetto sui dati.

Confronto grafico fra trasformazioni



Trasformazioni di tipo statistico

- Tra le principali trasformazioni di tipo statistico, segnaliamo:
  - ***differenze successive***
  - ***tassi di variazione***
  - ***centrare i dati***
  - ***standardizzazione***
  - ***riflessione***
  - ***normalizzazione***
  - . . . . . .

Differenze successive

- Queste trasformazioni hanno grande utilità soprattutto quando le osservazioni sono serie storiche x_t per $t = 1, 2, \ldots, n$.
- La ***differenza prima*** *rispetto al periodo precedente* è definita da:

$$\nabla x_t = x_t - x_{t-1}\,.$$

- La differenza prima esprime la variazione della variabile X_t intervenuta fra il tempo t e il precedente.
- La differenza prima su può re-iterare, applicando la differenza prima alla differenza prima; si definisce così la ***differenza seconda***:

$$\nabla^2 x_t = \nabla\nabla x_t = \nabla(x_t - x_{t-1}) = (x_t - x_{t-1}) - (x_t - x_{t-2}) = x_t - 2\,x_{t-1} + x_{t-2}.$$

Differenza stagionale

➤ Per dati che si ripetono s volte in un anno, la ***differenza stagionale*** è definita da:

$$\nabla_s x_t = x_t - x_{t-s} .$$

Così, $s = 12$ per dati mensili, $s = 6$ per dati bimestrali, $s = 4$ per dati trimestrali, etc. Per dati mensili, la differenza stagionale $\nabla_{12} x_t$ esprime la variazione intervenuta nella variabile X_t tra un mese e lo stesso mese dell'ano precedente.

➤ Si può applicare, sequenzialmente, la ***differenza prima alla differenza stagionale*** (ovvero, la differenza stagionale alla differenza prima) mediante la trasformazione:

$$\begin{aligned} \nabla \nabla_{12} x_t &= \nabla \left(x_t - x_{t-12} \right) = \left(x_t - x_{t-12} \right) - \left(x_{t-1} - x_{t-13} \right) \\ &= x_t - x_{t-1} - x_{t-12} + x_{t-13} . \end{aligned}$$

Effetti delle trasformazioni differenze

- La *differenza prima* ∇x_t elimina dalla serie originaria un trend lineare (cioè una variazione adattabile ad una linea retta).
- La *differenza stagionale* $\nabla_{12} x_t$ elimina dalla serie mensile originaria una componente stagionale ed un eventuale trend lineare.
- La *differenza prima della differenza stagionale* $\nabla\nabla_{12} x_t$ elimina dalla serie mensile originaria una componente stagionale ed un eventuale trend, anche di tipo quadratico.

Differenze sull'indice mensile della produzione industriale...1

➤ La Figura seguente presenta le $n = 168$ osservazioni della serie storica dell'Indice mensile della Produzione industriale (corretto per i giorni lavorativi) in Italia (Gennaio 1990 - Dicembre 2003), fonte ISTAT, assieme alla differenza prima, differenza stagionale e differenza prima della differenza stagionale.

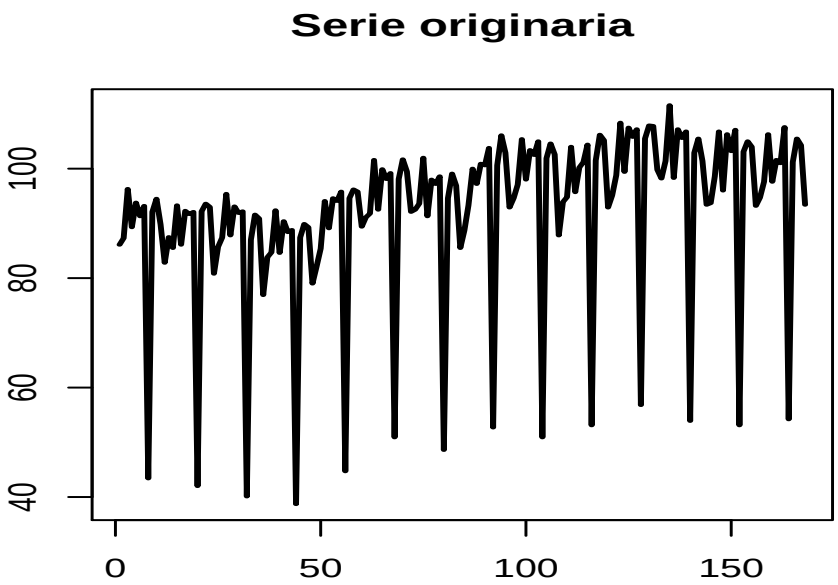
- La *serie originale* presenta una crescita lenta ed alternata fortemente caratterizzata dal crollo della produzione nel mese di Agosto per la interruzione quasi generalizzata dele attività produttve in Italia
- La *differenza prima* elimina quasi completamente la crescita nella serie e mostra un livello medio stabile e costante nel tempo; le due forti oscillazioni annuali sono dovute alla differenza Agosto-Luglio e Settembre-Agosto
- La *differenza stagionale* evidenzia un ciclo economico pluriennale nella serie e la quais completa eliminazione di un trend
- La *differenza della differenza stagionale* elimina sia la componente tendenziale che quella stagionale e mostra una sequenza di valori molto irregolare rispetto al tempo

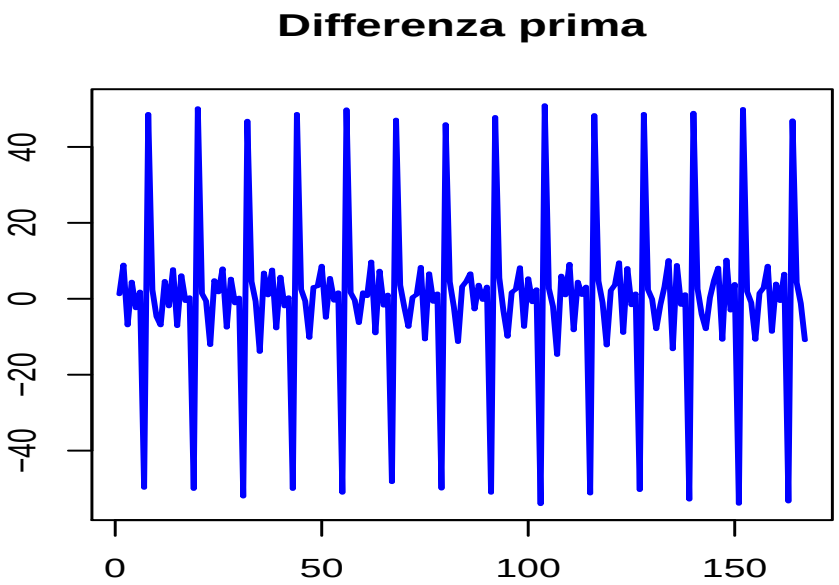
Differenze sull'indice mensile della produzione industriale...1

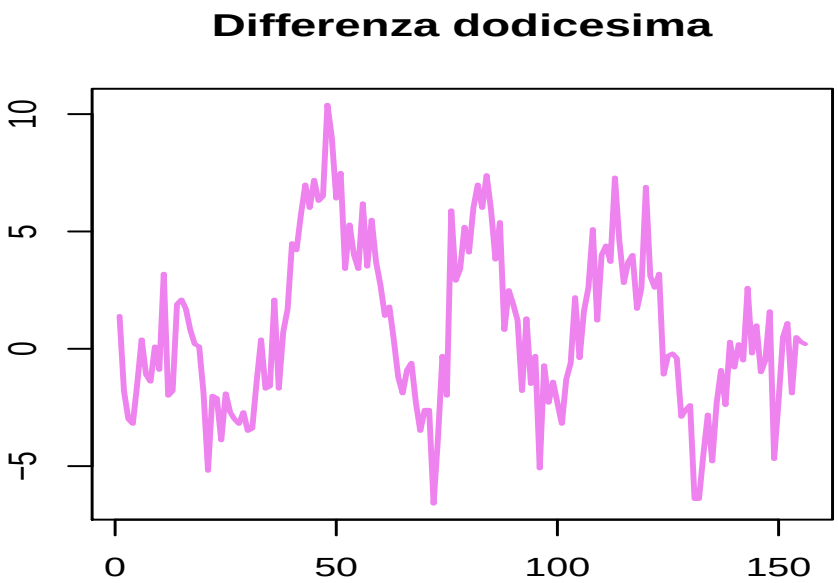
- Osservando i campi di variazione delle serie trasformate rispetto a quello della serie originaria (sull'asse delle ordinate), si nota come l'effetto di tutte le differenze è quello di ridurre in misura sensibile il campo di variazione della serie.

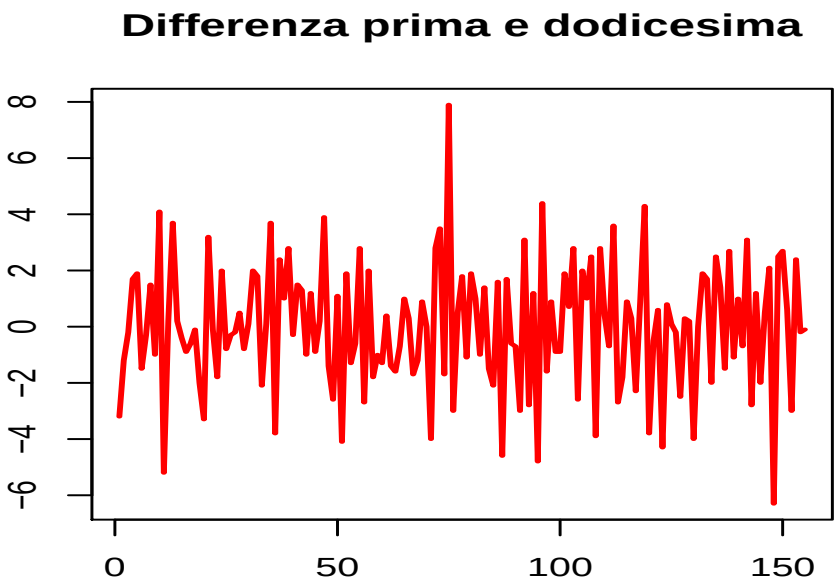
- Tali trasformazioni aiutano a comprendere i meccanismi che generano questi dati perché diminuiscono la variabilità della serie originaria.

Trasformazioni differenze sull'indice mensile della produzione industriale









Tassi di variazione

- Il ***tasso di variazione*** *rispetto al periodo precedente* è definito da:

$$Tv(x_t) = \frac{x_t - x_{t-1}}{x_{t-1}}.$$

- Il tasso di variazione non dipende dall'unità di misura della variabile e, talvolta, è moltiplicato per 100 (= tasso di variazione percentuale).

- Per dati che si ripetono s volte in un anno, il ***tasso di variazione*** *rispetto allo stesso periodo dell'anno precedente* è definito da:

$$Tv_s(x_t) = \frac{x_t - x_{t-s}}{x_{t-s}}.$$

Così, $s = 12$ per dati mensili, $s = 6$ per dati bimestrali, $s = 4$ per dati trimestrali, etc.

Variazione congiunturale e tendenziale

➤ Per dati mensili, e per i principali aggregati economici, il tasso di variazione rispetto al periodo precedente e il tasso di variazione rispetto allo stesso periodo dell'anno precedente vengono definiti, rispettivamente, ***variazione congiunturale*** e ***variazione tendenziale*** della serie.

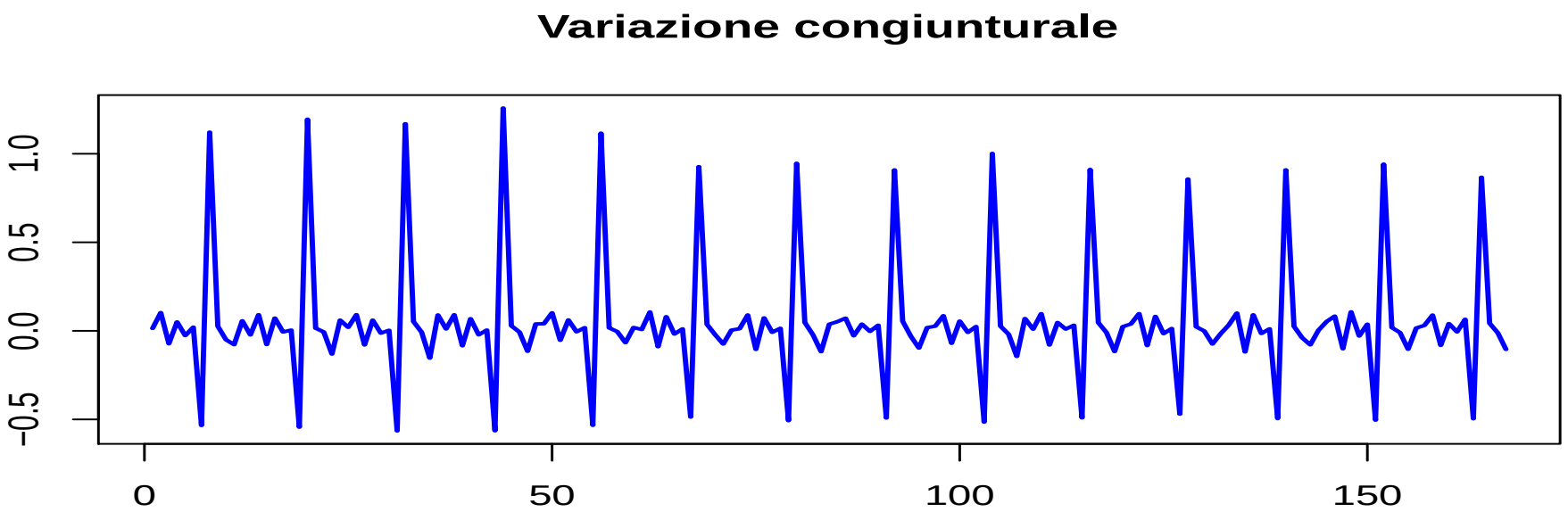
➤ Si dimostra che valgono le seguenti approssimazioni:

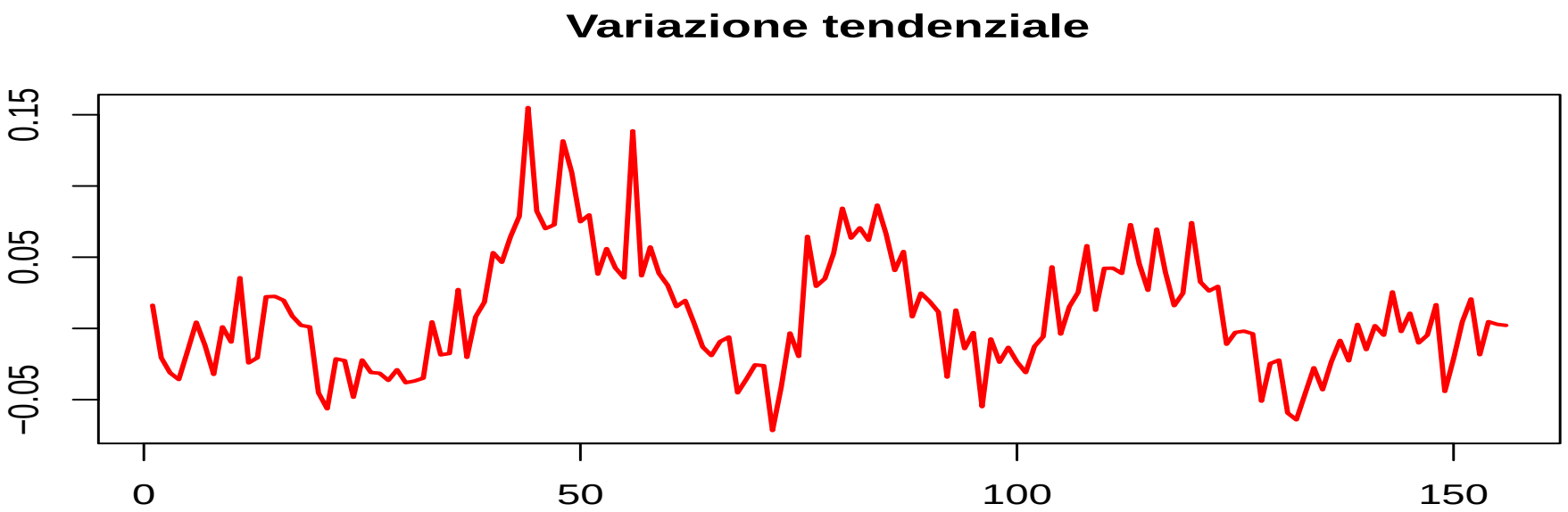
- **Variazione congiunturale:** $\nabla \log(x_t) \simeq \frac{x_t - x_{t-1}}{x_{t-1}} = Tv(x_t)$

- **Variazione tendenziale:** $\nabla_{12} \log(x_t) \simeq \frac{x_t - x_{t-12}}{x_{t-12}} = Tv_{12}(x_t)$

➤ La Figura successiva mostra, per l'Indice mensile della Produzione industriale in Italia, le serie storiche delle variazioni congiunturale e delle variazioni tendenziali.

Trasformazioni differenze sull'indice mensile della produzione industriale





➤ A fini di confronto tra variabili che si posizionano su livelli differenti, può essere utile centrare i dati mediante la trasformazione:

$$X^* = X - \mu\,.$$

In pratica, $X*$ rappresenta la variabile ***scarto*** dalla media (aritmetica): essa dipende dalla stessa unità di misura di X.

➤ Se si desiderano dati indipendenti dall'unità di misura, è conveniente utilizzare la ***standardizzazione*** della variabile X mediante:

$$X^* = \frac{X - \mu}{\sigma}\,.$$

➤ La *variabile standardizzata* ha media 0 e varianza 1.

Riflessione dei dati

- Quando si devono confrontare variabili che sono in ordine differente (per esempio, una crescente ed una decrescente) è opportuno operare trasformazioni che considerino tutte le variabili ordinate in modo omogeneo (per esempio, tutte crescenti).

- Ciò si verifica spesso per risposte che si esprimono mediante variabili ordinali, codificate per comodità con i numeri interi da 1 a m, così X assume le modalità $1, 2, \ldots, m$.

- Se ora si possiede una variabile X codificata all'incontrario, per cui il valore minimo è stato indicato con m ed il valore massimo è stato indicato con 1, si utilizza la riflessione della variabile X mediante la trasformazione:

$$X^* = m - X + 1$$

- In tal modo, i valori della X vengono ribaltati ("riflessi" come in uno specchio) ed il valore m diventa 1, il valore 1 diventa m e così via per tutti gli altri.

Normalizzazione dei dati

- Una variabile X che assume modalità quantitative sempre finite si può normalizzare mediante la trasformazione:

$$X^* = \frac{X - min(X)}{max(X) - min(X)} .$$

- La variabile X^* non dipenderà dall'unità di misura di X ed assumerà valori sempre compresi fra 0 e 1.

- La seguente trasformazione normalizzante

$$X^* = \frac{2\,X - max(X) - min(X)}{max(X) - min(X)} ,$$

genera una variabile X^* che assume sempre valori fra -1 e $+1$.

- Se invece la variabile X assume valori non-negativi, anche eventualmente infiniti, la seguente trasformazione normalizzante:

$$X^* = \frac{1}{1 + X}$$

assumerà sempre valori compresi fra 0 e 1.