Repubblica e Cantone Ticino
Dipartimeno delle finanze e dell'economia
Divisione delle risorse

Ufficio di statistica

SCOMPOSIZIONE DELLE DIFFERENZE SALARIALI TRA DUE GRUPPI

9

SCOMPOSIZIONE DELLE DIFFERENZE SALARIALI TRA DUE GRUPPI

Sandro Petrillo e Oscar Gonzalez

INDICE

4		**PREFAZIONE**
6	**1.**	**INTRODUZIONE**
8	**2.**	**IL QUADRO TEORICO**
9	2.1	Il metodo originale e il quadro generale
10	2.2	Assunzioni per identificare una distribuzione controfattuale e la scomposizione aggregata
18	**3.**	**PARTE EMPIRICA**
19	3.1	Introduzione
21	3.2	Esempio empirico
31		**APPENDICE A**
32		Codice R per replicare l'esempio empirico, con il pacchetto "decr"
37		**APPENDICE B**
38		Formule degli stimatori utilizzati
38	B.1	Stimatore degli effettivi (numero di salariati o di posti di lavoro)
38	B.2	Funzione di ripartizione empirica dei salari
38	B.3	Quantili dei salari (mediana, percentili, ...)
38	B.4	Salario medio
40		**APPENDICE C**
41		R session info
42		**BIBLIOGRAFIA**

PREFAZIONE

Il principale compito della statistica pubblica è quello di creare e diffondere informazioni statistiche, offrendo strumenti fondamentali a tutti i cittadini per conoscere e analizzare la società e le sue evoluzioni e, in definitiva, farsi un'opinione. Questo rende la statistica pubblica uno dei pilastri del processo democratico. Questo dovrebbe inoltre favorire lo sviluppo di idee e la ricerca di soluzioni ai problemi della società e, in definitiva, costituire un rilevante supporto ai processi decisionali.

Nell'adempimento della sua missione, la statistica pubblica si caratterizza attraverso l'adozione di principi scientifici e deontologici forti, così come definiti nella Legge sulla statistica cantonale (LStaC) e nella Carta della statistica pubblica svizzera. Questo rigore costituisce un presupposto fondamentale per poter fornire "informazioni pertinenti, corrette e imparziali" (art.3 cpv.1 LStaC), così come prevede la legge.

Uno dei principi che caratterizzano le attività della statistica pubblica è quello della trasparenza. Il quarto principio della Carta della statistica pubblica svizzera recita infatti che "le informazioni statistiche sono documentate in modo tale da agevolarne la comprensione e da consentirne un uso corretto". In effetti, quanto detto fin qui non basta: la pertinenza, la correttezza e l'oggettività delle informazioni devono poter essere verificate da chiunque, e questo è possibile solo se la statistica mette a disposizione le necessarie informazioni.

Questo documento rappresenta un buon esempio di concretizzazione di quanto fin qui esposto: il tema affrontato è quello dei salari, più specificatamente del confronto tra salari di gruppi diversi di lavoratori. In effetti, in questo specifico ambito tematico il fabbisogno informativo dell'utente è quasi sempre quello di mettere a confronto le retribuzioni di due o più gruppi di individui. Il problema è che un confronto diretto dei salari cela qualche insidia, perché i gruppi – sebbene sempre accomunati da un tratto specifico, che li definisce (appunto) come gruppi – possono avere una certa eterogeneità interna. In questo senso, le differenze salariali osservate possono essere sì da ricondurre a delle differenze retributive, ma non solo, perché queste possono derivare anche da delle differenze nelle caratteristiche strutturali dei lavoratori, per esempio in termini di profili formativi, di impiego in rami economici o imprese a diversa retribuzione, ecc. Queste differenze hanno un effetto negativo sul confronto, che ne risulta distorto, e che può portare l'analista a trarre considerazioni errate sull'entità e la natura delle disparità salariali misurate.

Tenendo conto dell'importanza e della sensibilità del tema, così come della difficoltà empirica nel condurre questo tipo di analisi, questo contributo si propone di fornire all'utenza (piuttosto esperta) una metodologia per scomporre le differenze che si osservano tra le statistiche dei salari di due gruppi di individui e, al contempo, mettere a disposizione un pacchetto per il software R (linguaggio di programmazione e ambiente di sviluppo specifico per l'analisi statistica dei dati) denominato *decr* e sviluppato dall'Ustat per l'implementazione empirica di tale metodologia.

Pau Origoni
Capoufficio, Ufficio di statistica

1. INTRODUZIONE

In questo documento intendiamo presentare un metodo per scomporre le differenze che si osservano tra le statistiche dei salari di due gruppi di individui e un pacchetto R, denominato "decr", per implementare empiricamente tali scomposizioni.

I due gruppi di individui sui quali si vogliono esaminare le differenze salariali possono essere, per esempio, uomini e donne in un determinato momento del tempo, residenti e frontalieri, oppure, i salariati a una certa data e i salariati a un'altra data.

Per inquadrare meglio il problema, partiamo da qualche esempio di statistiche estratte dalla Rilevazione svizzera della struttura dei salari (RSS).

Secondo i dati della RSS, nel 2014 in Ticino nell'economia privata il salario mensile lordo standardizzato mediano degli uomini era di 5.397 franchi, mentre quello delle donne di 4.546 franchi. Si osserva quindi, in termini di mediana, una differenza di salario di 851 franchi a sfavore delle donne.

Prendendo come esempio i salariati svizzeri e quelli frontalieri, le statistiche della RSS ci dicono che in Ticino il salario mensile mediano dei primi, nel 2014, era di 5.694 franchi, mentre quello dei secondi di 4.523 franchi. Il salario mediano degli svizzeri risulta quindi essere superiore di 1.171 franchi rispetto a quello dei frontalieri.

Gli esempi appena elencati riflettono quello che si osserva sul mercato del lavoro, senza tener conto di eventuali differenze nelle caratteristiche dei due gruppi di individui (uomini e donne nel primo esempio, svizzeri e frontalieri nel secondo). L'obiettivo della scomposizione tra due gruppi è proprio quello di suddividere la differenza osservata, in termini di una statistica distribuzionale (negli esempi la mediana), in due componenti:

- Una componente spiegabile, attribuibile o giustificabile dal fatto che i due gruppi hanno caratteristiche diverse. Questa componente viene anche chiamata "parte spiegata";
- Una componente strutturale, che riflette eventuali differenze nella struttura dei salari dei due gruppi. In altri termini, questa componente, se diversa da zero, ci direbbe che i due gruppi, a parità di caratteristiche, sono remunerati in maniera diversa. Questa componente viene anche chiamata "parte non spiegata".

Questo documento si articola in tre capitoli. Dopo un primo capitolo introduttivo, il secondo espone il quadro teorico, in cui nella prima sezione vengono presentati il metodo originale e il quadro generale che definisce le differenze tra le statistiche salariali di due gruppi di individui, mentre nella seconda sezione vengono mostrate le assunzioni necessarie all'identificazione di una distribuzione dei salari controfattuale e le scomposizioni delle differenze tra le statistiche salariali di due gruppi di individui.

Il terzo capitolo, di stampo empirico, illustra in primo luogo lo stimatore del fattore di riponderazione, che serve per la stima della distribuzione controfattuale vista nel secondo capitolo, e in secondo luogo un esempio empirico con dei dati fittizi. L'esempio empirico mostra passo per passo un'applicazione del metodo proposto in questo documento per la scomposizione delle differenze tra le statistiche salariali di due gruppi di individui.

2.
IL QUADRO TEORICO

2.1 Il metodo originale e il quadro generale

Il metodo originale, proposto dagli articoli di Blinder (1973) e Oaxaca (1973) (che chiameremo BO da qui in avanti), consisteva nella scomposizione dei salari medi tra due gruppi nelle due componenti (spiegabile e strutturale, o spiegata e non spiegata), ipotizzando che i salari (Y) fossero determinati linearmente da delle caratteristiche osservate (X) e che il termine di errore ε_i fosse condizionatamente indipendente da X:

$$Y_{gi} = \beta_{g0} + \sum_{k=1}^{K} X_{ik}\beta_{gk} + \varepsilon_{gi}, \quad g = A, B, \tag{2.1}$$

dove il valore atteso del termine di errore, condizionato alle caratteristiche osservate, è uguale a zero ($E(\varepsilon_{gi}|X_i) = 0$) e X_i è il vettore delle K caratteristiche osservate dell'individuo i ($X_i = [X_{i1}, \ldots, X_{iK}]$). In questo contesto, vengono stimate due equazioni di regressione (una per gruppo), con il metodo dei minimi quadrati ordinari, che permettono di esprimere la differenza osservata tra le medie dei salari dei due gruppi nella maniera seguente:

$$\begin{aligned}\widehat{\Delta}_O^{\mu} &= \widehat{\overline{Y}}_A - \widehat{\overline{Y}}_B = \overline{X}'_A\widehat{\beta}_A - \overline{X}'_B\widehat{\beta}_B = \\ &= \underbrace{(\widehat{\beta}_{A0} - \widehat{\beta}_{B0}) + \sum_{k=1}^{K} \overline{X}_{Bk}\left(\widehat{\beta}_{Ak} - \widehat{\beta}_{Bk}\right)}_{\widehat{\Delta}_S^{\mu}\text{ (Parte non spiegata)}} + \underbrace{\sum_{k=1}^{K}(\overline{X}_{Ak} - \overline{X}_{Bk})\widehat{\beta}_{Ak}}_{\widehat{\Delta}_X^{\mu}\text{ (Parte spiegata)}},\end{aligned}$$

dove $\widehat{\beta}_{g0}$ e $\widehat{\beta}_{gk}$ $(k = 1, \ldots, K)$ sono i coefficienti stimati dei modelli di regressione (lineare) per i gruppi $g = A, B$.

Negli esempi dell'introduzione sono state elencate delle differenze osservate tra i salari mediani di due gruppi, e il metodo originale BO esposto brevemente prevede la scomposizione della differenza tra i salari medi di due gruppi. La mediana e la media sono solo due delle molte statistiche che si possono esprimere in termini della funzione di distribuzione cumulata, o funzione di ripartizione, dei salari. Indicando con Y i salari, la funzione di distribuzione cumulata è, per definizione:

$$F_Y(y) = Pr(Y \leq y).$$

La funzione di ripartizione, $F_Y(y)$, indica la probabilità, compresa tra zero e uno, che i salari Y siano inferiori o uguali a un determinato valore y.

Ci focalizziamo sulla funzione di ripartizione (dei salari), dato che serve da base con cui si possono esprimere varie statistiche. In particolare, la utilizziamo per formulare una statistica distribuzionale "generica" con $\nu(F_Y)$, dove $\nu : \mathcal{F}_\nu \to \mathbb{R}$ è una funzione reale e $\mathcal{F}_\nu$ è una classe di funzioni di distribuzione tali $F_Y \in \mathcal{F}_\nu$ che se $|\nu(F_Y)| < \infty$.

Le differenze di salario considerate in questo documento riguardano unicamente due gruppi mutualmente esclusivi, che indichiamo con A e B. Quindi, per un individuo i della popolazione in esame, $D_{Ai} + D_{Bi} = 1$, dove $D_{gi} = \mathbf{1}\{i \text{ appartiene a } g\}$, $g = A, B$ e $\mathbf{1}\{\cdot\}$ è la funzione indicatrice.

In questo contesto, la differenza di salario tra due gruppi, A e B, in termini di una statistica distribuzionale ν, può essere espressa nel modo seguente:

$$\Delta_O^{\nu} = \nu(F_{Y_A|D_A}) - \nu(F_{Y_B|D_B}), \tag{2.2}$$

dove $F_{Y_A|D_A}$ e $F_{Y_B|D_B}$ sono le funzioni di ripartizione dei salari degli individui dei gruppi A e B.

La notazione $Y_A|D_A$, ripresa da Fortin, Lemieux e Firpo (2011), viene utilizzata per sottolineare il fatto che la distribuzione dei salari di un gruppo (Y_A) è "implicitamente" condizionata all'appartenenza di un individuo al gruppo A ($|D_A$). Più in generale, possiamo esprimere una funzione di ripartizione dei salari con $F_{Y_g|D_s}$, dove sia g che s possono prendere i valori A e B. Nel caso in cui $g = s$, $F_{Y_g|D_s}$ è una distribuzione dei salari osservata ($F_{Y_A|D_A}$) e ($F_{Y_B|D_B}$). Invece, nel caso in cui $g \neq s$, $F_{Y_g|D_s}$ è una distribuzione **controfattuale** (non osservata). A grandi linee, la distribuzione controfattuale $F_{Y_A|D_B}$ dovrebbe rappresentare la distribuzione dei salari degli individui del gruppo B se questi fossero remunerati secondo la struttura dei salari del gruppo A. Si tratta della quantità fondamentale che permetterebbe di scomporre la differenza dei salari osservata tra due gruppi in una componente "spiegata" e una "non spiegata".

Dato che, per definizione, la distribuzione dei salari controfattuale non è osservata, è necessario introdurre qualche assunzione per poterla identificare. Nella prossima sezione presentiamo alcune assunzioni necessarie a identificare, nei dati a diposizione, un tipo di distribuzione controfattuale dei salari. L'identificazione di una distribuzione controfattuale permetterà di riflesso di identificare la statistica controfattuale ν, che a sua volta sarà la quantità chiave per poter effettuare la scomposizione delle differenze salariali osservate.

Una volta presentate le quantità da stimare e le loro condizioni di identificazione, nella sezione successiva presenteremo un metodo empirico per effettuare la scomposizione delle differenze salariali (applicabile a qualsiasi statistica che si può esprimere attraverso la funzione di ripartizione dei salari).

2.2 Assunzioni per identificare una distribuzione controfattuale e la scomposizione aggregata

In questa sezione utilizzeremo, come nella sezione precedente, la notazione utilizzata nel documento di Fortin, Lemieux e Firpo (2011). Come accennato nell'introduzione, l'obiettivo della scomposizione delle differenze salariali tra due gruppi è quello di dividere Δ_O^ν in una componente attribuibile alle differenze dei due gruppi in termini di caratteristiche osservate (Δ_X^ν, parte spiegata) e una componente dovuta a delle differenze nelle strutture dei salari (Δ_S^ν, parte non spiegata). La scomposizione della differenza osservata in queste due componenti dipende dalla costruzione appropriata di una distribuzione dei salari **controfattuale**. Nel contesto delle scomposizioni delle differenze salariali, i controfattuali utilizzati consistono nella rielaborazione delle strutture salariali, che collegano le caratteristiche osservate e non nei due gruppi ai loro salari. Per facilitare questo legame, il ruolo della struttura salariale viene formalizzato nell'assunzione seguente:

ASSUNZIONE 1 [**Forma strutturale**] Un lavoratore i, che appartiene a uno dei due gruppi A o B, è remunerato secondo una delle strutture dei salari, m_A o m_B, che sono funzioni delle caratteristiche osservabili (X) e non osservabili (ε) del lavoratore:

$$Y_{Ai} = m_A(X_i, \varepsilon_i) \; e \; Y_{Bi} = m_B(X_i, \varepsilon_i), \tag{2.3}$$

dove ε_i ha una distribuzione condizionata $F_{\varepsilon|X}$ dato X.

Questa assunzione non pone restrizioni in termini di forme funzionali delle strutture dei salari. Il suo contributo principale è di formalizzare il fatto che i fattori alla base della determinazione del salario di un individuo possono essere da una parte le sue caratteristiche osservate (X_i) e dall'altra anche le sue caratteristiche non osservate ε_i. Ciò che implica questa assunzione è che ci sono solo tre ragioni per le quali la distribuzione dei salari dei gruppi A e B può differire. Le tre potenziali fonti di differenza sono:

- differenze tra le funzioni di determinazione del salario, m_A e m_B,
- differenze nelle distribuzioni delle caratteristiche osservabili (X), e
- differenze nelle distribuzioni delle caratteristiche non osservabili (ε).

L'obiettivo della scomposizione delle differenze salariali tra i due gruppi, che per semplicità chiameremo scomposizione aggregata da qui in avanti, è quello di separare il contributo del primo fattore (differenze tra le strutture salariali m_A e m_B) dagli altri due.

La prossima assunzione viene introdotta per delimitare il tipo di controfattuale che verrà utilizzato per la scomposizione aggregata.

ASSUNZIONE 2 [**Controfattuale di tipo semplice**] Una struttura di salario controfattuale, m^C, corrisponde a un controfattuale di tipo semplice quando si può assumere che $m^C(\cdot,\cdot) \equiv m_A(\cdot,\cdot)$ per lavoratori del gruppo B, o $m^C(\cdot,\cdot) \equiv m_B(\cdot,\cdot)$ per lavoratori del gruppo A.

Teoricamente, nel contesto del paragone dei salari di due gruppi, ci si potrebbe interessare a un tipo di controfattuale che risponderebbe alla domanda seguente:

- Come sarebbe la distribuzione dei salari del gruppo A se questi fossero remunerati secondo la struttura dei salari che esisterebbe in assenza dei salariati del gruppo B sul mercato del lavoro?

Ciò che implica l'assunzione 2 è proprio quello di escludere questo tipo di controfattuali, restringendo il campo a qualcosa di più realizzabile con i dati a disposizione. L'assunzione 2 equivale a escludere i cosiddetti "effetti di equilibrio generale".

Si tratta di un primo passo verso l'identificazione di una distribuzione dei salari controfattuale.

In pratica questa assunzione ci permette di utilizzare le osservazioni di un gruppo (quindi la loro struttura remunerativa) con le caratteristiche dell'altro. Questo dovrebbe portare in seguito al paragone di individui con caratteristiche simili che differirebbero "solo" nelle loro strutture di salario. Per garantire il paragone di individui con caratteristiche simili, è necessario che non esistano individui di un gruppo con caratteristiche che nessun individuo dell'altro gruppo abbia. In altri termini, la scomposizione aggregata potrà essere fatta unicamente nel "supporto comune", ovvero in quell'insieme di combinazioni di caratteristiche per le quali esiste almeno un individuo di entrambi i gruppi. Questo viene formalizzato nella prossima assunzione:

ASSUNZIONE 3 [**Sovrapposizione del supporto**] Si consideri $\mathcal{X} \times \mathcal{E}$ come il supporto di tutti i fattori che determinano i salari $[X', \varepsilon']'$. Per ogni $[x', e']'$ appartenenti all'insieme $\mathcal{X} \times \mathcal{E}$, $0 < Pr[D_B = 1|X = x, \varepsilon = e] < 1$.

Questa assunzione è cruciale: come già accennato, rende attenti al non paragonare individui non paragonabili perché (troppo) diversi rispetto alle loro caratteristiche. Un altro segnale che dà questa assunzione è che la scomposizione aggregata non riguarda per forza la differenza di salario osservata, in particolare nel caso in cui le caratteristiche dei due gruppi in esame non siano (in parte) le stesse. La scomposizione nelle due componenti è possibile e va fatta unicamente nella zona del supporto comune tra i due gruppi. In altri termini, la differenza Δ_O^ν non corrisponde per forza alla differenza di salario osservata tra i due gruppi **nel supporto comune**, ossia in quell'insieme di caratteristiche per le quali esistono individui di entrambi i gruppi. Nel caso della scomposizione della differenza dei salari medi di due gruppi, Nopo (2008) ha proposto una scomposizione alternativa che integra anche le eventuali differenze nei supporti delle caratteristiche dei due gruppi che, come vedremo in seguito, possono essere considerate come una parte spiegata delle differenze salariali osservate (spiegata dal fatto che i due gruppi hanno dei profili non paragonabili in termini di caratteristiche). Nel caso di altre statistiche distribuzionali, come la mediana e altri quantili, questa scomposizione alternativa non è possibile e, quindi, la scomposizione va fatta solo nel supporto comune ai due gruppi, indicando possibilmente, a parte, quali sono le caratteristiche dei due eventuali sottoinsiemi di individui fuori dal supporto comune (sia in termini di caratteristiche che di distribuzione dei salari). Contrariamente al caso della media, eventuali differenze salariali dovute a delle differenze nei supporti delle caratteristiche dei due gruppi non sono scomponibili in due componenti da attribuire ai due gruppi. Un'eventuale differenza "residua" (tra la differenza osservata e quella nel supporto comune) può comunque essere considerata e aggiunta alla parte spiegata che risulta dalla scomposizione nel supporto comune. Entreremo nei dettagli della scomposizione alternativa di Nopo (2008) in seguito, nella parte del metodo empirico di questo documento, per il caso specifico del salario medio.

A questo punto, avendo stabilito il tipo di controfattuale che utilizzeremo (assunzione 2) e fatto l'assunzione della sovrapposizione del supporto (assunzione 3), possiamo presentare il modo con cui verrà creata la distribuzione dei salari controfattuale.

Per ogni membro dei due gruppi $g = A, B$, si osservano una variabile di interesse Y_{gi} (il salario) e alcune caratteristiche individuali X_i. Y_g e X hanno una distribuzione congiunta condizionata, $F_{Y_g,X|D_g}(\cdot,\cdot) : \mathbb{R} \times \mathcal{X} \to [0,1]$, e $\mathcal{X} \subset \mathbb{R}^k$ è il supporto di X. Considerando le distribuzioni delle caratteristiche (X) nei due gruppi, $X|D_A$ e $X|D_B$, si possono suddividere le caratteristiche osservate in due gruppi:

- il supporto comune, $\mathcal{S}$, che comprende tutte le combinazioni di caratteristiche per le quali si osservano individui di entrambi i gruppi,
- e il suo complemento, $\mathcal{S}^C$, che comprende tutte le combinazioni di caratteristiche per le quali si osservano o solo individui del gruppo A o solo individui del gruppo B.

Il supporto comune $\mathcal{S}$ (e il suo complemento $\mathcal{S}^C$) e i due gruppi ($g = A, B$) permettono di suddividere la popolazione in esame in quattro sottogruppi:

1. $g = A, X \in \mathcal{S}$: individui del gruppo A con caratteristiche del supporto comune,
2. $g = B, X \in \mathcal{S}$: individui del gruppo B con caratteristiche del supporto comune,

3. $g = A, X \in \mathcal{S}^C$: individui del gruppo A con caratteristiche che non appartengono al supporto comune,
4. $g = B, X \in \mathcal{S}^C$: individui del gruppo B con caratteristiche che non appartengono al supporto comune.

La differenza salariale tra i due gruppi che si andrà a scomporre riguarda i primi due sottogruppi. A questo proposito definiamo le distribuzioni salariali di questi due sottogruppi:

$$Pr(Y \leq y | g = A, X \in \mathcal{S}) = F_{Y|D_A,\mathcal{S}}(y) = F_{Y_A|\mathcal{S}}(y) = F_{Y_A|D_A,\mathcal{S}}(y)$$

e

$$Pr(Y \leq y | g = B, X \in \mathcal{S}) = F_{Y|D_B,\mathcal{S}}(y) = F_{Y_B|\mathcal{S}}(y) = F_{Y_B|D_B,\mathcal{S}}(y),$$

per poi presentare la differenza tra i due gruppi, in termini di una statistica distribuzionale, osservata all'interno del supporto comune:

$$\Delta^{\nu}_{O|\mathcal{S}} = \nu(F_{Y_A|D_A,\mathcal{S}}) - \nu(F_{Y_B|D_B,\mathcal{S}}). \tag{2.4}$$

Si noti come l'equazione (2.4) sia diversa dalla (2.2). La (2.2) misura la differenza tra le statistiche salariali di due gruppi che si osserva considerando tutti gli individui della popolazione, mentre la (2.4) considera solo gli individui dei due gruppi che fanno parte del supporto comune, ovvero che hanno caratteristiche paragonabili. È quest'ultima differenza che può essere scomposta in una parte spiegata, dalla diversa distribuzione delle caratteristiche, e in una non spiegata. Nel supporto comune, infatti, nonostante i due gruppi abbiano lo stesso ventaglio di caratteristiche, queste possono essere distribuite in maniera differente. L'obiettivo della scomposizione è proprio quello di creare una distribuzione dei salari "a parità di distribuzione di caratteristiche", che permette di separare l'effetto della differenza di caratteristiche da quello della differenza nelle strutture salariali.

La distribuzione di $Y_g|D_g, \mathcal{S}$ viene definita utilizzando la legge delle aspettative iterate[1] (law of iterated expectations), e cioè, una volta che si integra sulle caratteristiche osservate, si ottiene:

$$F_{Y_g|D_g,\mathcal{S}}(y) = \int F_{Y_g|X,D_g,\mathcal{S}}(y|X = x) \cdot dF_{X|D_g,\mathcal{S}}(x), \quad g = A, B. \tag{2.5}$$

[1] La legge delle aspettative iterate può essere applicata a una funzione di ripartizione di una variabile aleatoria Y, considerando che può essere espressa come il valore atteso di una funzione indicatrice: $F_Y(y) = Pr(Y \leq y) = Pr(Y \in]-\infty; y]) = E(\mathbf{1}_{]-\infty;y]}\{Y\})$.

Utilizzando l'equazione (2.5) è possibile esprimere la distribuzione dei salari dei due gruppi A e B nel supporto comune:

$$F_{Y_A|D_A,\mathcal{S}}(y) = \int F_{Y_A|X,D_A,\mathcal{S}}(y|X = x) \cdot dF_{X|D_A,\mathcal{S}}(x),$$

e

$$F_{Y_B|D_B,\mathcal{S}}(y) = \int F_{Y_B|X,D_B,\mathcal{S}}(y|X = x) \cdot dF_{X|D_B,\mathcal{S}}(x).$$

L'equazione (2.5) può essere usata anche per costruire una distribuzione controfattuale dei salari, mischiando la distribuzione condizionata dei salari del gruppo A date le caratteristiche X nel gruppo A ($F_{Y_A|X,D_A,\mathcal{S}}$) con la distribuzione delle caratteristiche del gruppo B ($F_{X|D_B,\mathcal{S}}$), ottenendo:

$$F_{Y^C_A:X=X|D_B,\mathcal{S}}(y) = \int F_{Y_A|X,D_A,\mathcal{S}}(y|X = x) \cdot dF_{X|D_B,\mathcal{S}}(x). \tag{2.6}$$

La distribuzione controfattuale dei salari formulata nell'equazione (2.6) rappresenta la distribuzione dei salari degli individui del gruppo B se questi fossero remunerati secondo la struttura salariale dei salariati del gruppo A. Secondo una lettura in un'altra ottica, rappresenta anche la distribuzione dei salari dei lavoratori del gruppo A se le loro caratteristiche (X) fossero distribuite come quelle del gruppo B. Si noti che la C in apice ($F_{Y_A^C}$) è stata aggiunta per sottolineare che si tratta di una distribuzione controfattuale dei salari.

La distribuzione controfattuale dei salari è stata costruita partendo dalla distribuzione dei salari del gruppo B, sostituendo la loro distribuzione condizionata dei salari ($F_{Y_B|X,D_B,\mathcal{S}}$) con la distribuzione condizionata dei salari del gruppo A ($F_{Y_A|X,D_A,\mathcal{S}}$). Esprimendo questa distribuzione condizionata attraverso la forma strutturale dei salari,

$$F_{Y_A|X,D_A,\mathcal{S}}(y|X=x) = Pr(Y_A \leq y|X=x, D_A=1, \mathcal{S}) = Pr(m_A(X,\varepsilon) \leq y|X=x, D_A=1, \mathcal{S}),$$

si vede come la distribuzione dei salari del gruppo A, condizionata alle caratteristiche osservate X, dipende dalla struttura dei salari, $m_A(\cdot)$, e dalla distribuzione condizionata delle caratteristiche non osservate, $\varepsilon|X$.

Nella costruzione della distribuzione controfattuale dell'equazione (2.6), avendo sostituito la distribuzione condizionata dei salari del gruppo B ($F_{Y_B|X,D_B,\mathcal{S}}$) con la distribuzione condizionata dei salari del gruppo A ($F_{Y_A|X,D_A,\mathcal{S}}$), sono state sostituite sia la struttura salariale che la distribuzione condizionata delle caratteristiche non osservate ε. Quindi, a meno che non vengano imposte delle assunzioni sulla distribuzione condizionata di ε, la distribuzione controfattuale costruita non porterebbe a dei risultati interpretabili dato che mescolerebbe delle differenze nelle strutture salariali e nella distribuzione di ε. Per esempio, a meno che le caratteristiche non osservate (ε) abbiano la stessa distribuzione condizionata (alle caratteristiche osservate X) nei due gruppi, la differenza tra le distribuzioni salariali

$$\begin{aligned}
&F_{Y_A^C:X=X|D_B,\mathcal{S}}(y) - F_{Y_B|D_B,\mathcal{S}}(y) = \\
&= \int F_{Y_A|X,D_A,\mathcal{S}}(y|X=x) \cdot dF_{X|D_B,\mathcal{S}}(x) - \int F_{Y_B|X,D_B,\mathcal{S}}(y|X=x) \cdot dF_{X|D_B,\mathcal{S}}(x) = \\
&= \int \left(F_{Y_A|X,D_A,\mathcal{S}}(y|X=x) - F_{Y_B|X,D_B,\mathcal{S}}(y|X=x)\right) \cdot dF_{X|D_B,\mathcal{S}}(x) = \\
&= \int \left(Pr(Y \leq y|X=x, D_A=1, \mathcal{S}) - Pr(Y \leq y|X=x, D_B=1, \mathcal{S})\right) \cdot \mathrm{d}F_{X|D_B,\mathcal{S}}(x) = \\
&= \int \left(Pr(m_A(X,\varepsilon) \leq y|X=x, D_A=1, \mathcal{S}) - Pr(m_B(X,\varepsilon) \leq y|X=x, D_B=1, \mathcal{S})\right) \cdot \mathrm{d}F_{X|D_B,\mathcal{S}}(x),
\end{aligned} \tag{2.7}$$

potrebbe riflettere non solo delle differenze tra le strutture salariali dei due gruppi, m_A e m_B, ma anche delle differenze tra le distribuzioni condizionate delle caratteristiche non osservate nei due gruppi, $\varepsilon|X$. Per fare in modo che questo termine rifletta unicamente differenze tra le strutture salariali dei due gruppi, m_A e m_B, è necessario che le distribuzioni condizionate delle caratteristiche non osservate, $\varepsilon|X$, siano le stesse nei due gruppi. Per questo motivo, viene inserita la prossima assunzione:

ASSUNZIONE 4 [**Indipendenza condizionata**] Per $g = A, B$, si consideri che (D_g, X, ε) abbiano una distribuzione congiunta. Per ogni x appartenente a $\mathcal{X}$: ε è indipendente da D_g dato $X = x$ o, equivalentemente, $D_g \perp \varepsilon|X$.

Grazie a quest'ultima assunzione, si arriva al principale risultato che permette l'identificazione della scomposizione aggregata:

PROPOSIZIONE 1 [**Identificazione della scomposizione aggregata**]

Sulla base delle assunzioni 2 (controfattuale semplice), 3 (sovrapposizione del supporto), e 4 (indipendenza condizionata), la differenza tra i salari dei gruppi A e B nel supporto comune $\mathcal{S}$, misurata in termini di una statistica distribuzionale ν ($\Delta^{\nu}_{O|\mathcal{S}}$), può essere scritta in questo modo:

$$\begin{aligned}\Delta^{\nu}_{O|\mathcal{S}} &= \nu(F_{Y_A|D_A,\mathcal{S}}) - \nu(F_{Y_B|D_B,\mathcal{S}}) = \\ &= \Big(\nu(F_{Y_A|D_A,\mathcal{S}}) - \nu(F_{Y^C_A:X=X|D_B,\mathcal{S}})\Big) + \Big(\nu(F_{Y^C_A:X=X|D_B,\mathcal{S}}) - \nu(F_{Y_B|D_B,\mathcal{S}})\Big) = \\ &= \Delta^{\nu}_{X} + \Delta^{\nu}_{S},\end{aligned} \tag{2.8}$$

dove

- il termine dell'effetto di composizione $\Delta^{\nu}_{X} = \nu(F_{Y_A|D_A,\mathcal{S}}) - \nu(F_{Y^C_A:X=X|D_B,\mathcal{S}})$ riflette unicamente l'effetto delle differenze nelle distribuzioni delle caratteristiche osservate (X) tra i due gruppi;
- il termine della stuttura salariale $\Delta^{\nu}_{S} = \nu(F_{Y^C_A:X=X|D_B,\mathcal{S}}) - \nu(F_{Y_B|D_B,\mathcal{S}})$ riflette unicamente differenze tra le funzioni strutturali $m_A(\cdot,\cdot)$ e $m_B(\cdot,\cdot)$.

Come abbiamo visto nell'equazione (2.7), sulla base delle assunzioni 3 (sovrapposizione del supporto) e 4 (indipendenza condizionata), l'unica fonte di differenza tra le distribuzioni dei salari $F_{Y^C_A:X=X|D_B,\mathcal{S}}$ e $F_{Y_B|D_B,\mathcal{S}}$ è la differenza tra le strutture salariali dei due gruppi, $m_A(\cdot)$ e $m_B(\cdot)$. Sulla base delle assunzioni 3 e 4 abbiamo $\Delta^{\nu}_{O|\mathcal{S}} = \nu(F_{Y_A|D_A,\mathcal{S}}) - \nu(F_{Y^C_A:X=X|D_B,\mathcal{S}}) + \Delta^{\nu}_{S}$, dove

$$\begin{aligned}F_{Y_A|D_A,\mathcal{S}}(y) - F_{Y^C_A:X=X|D_B,\mathcal{S}}(y) = \\ &= \int F_{Y_A|X,D_A,\mathcal{S}}(y|X=x)\cdot dF_{X|D_A,\mathcal{S}}(x) - \int F_{Y_A|X,D_A,\mathcal{S}}(y|X=x)\cdot dF_{X|D_B,\mathcal{S}}(x) = \\ &= \int F_{Y_A|X,D_A,\mathcal{S}}(y|X=x)\cdot \big(dF_{X|D_A,\mathcal{S}}(x) - dF_{X|D_B,\mathcal{S}}(x)\big) = \\ &= \int Pr(Y\le y|X=x,D_A=1,\mathcal{S})\cdot \big(\mathrm{d}F_{X|D_A,\mathcal{S}}(x) - \mathrm{d}F_{X|D_B,\mathcal{S}}(x)\big)\,.\end{aligned}$$

Quindi, il termine della scomposione aggregata $\Delta^{\nu}_{X} = \nu(F_{Y_A|D_A,\mathcal{S}}) - \nu(F_{Y^C_A:X=X|D_B,\mathcal{S}})$ è dovuto unicamente a delle differenze tra le distribuzioni delle caratteristiche osservate, X, nei due gruppi.

Una caratteristica vantaggiosa della scomposizione aggregata è che può essere eseguita senza nessuna assunzione sulle forme funzionali delle strutture salariali, $m_g(X,\varepsilon)$, pur vincolando, parzialmente, la distribuzione delle caratteristiche non osservate (ε). Le differenze nelle distribuzioni delle caratteristiche non osservate sono vincolate "ragionevolmente" attraverso l'assunzione di indipendenza condizionata. Questa assunzione, infatti, non esclude che le distribuzioni delle caratteristiche non osservate siano differenti nei due gruppi, ma assume che le distribuzioni delle caratteristiche non osservate, condizionate alle caratteristiche osservate (X), siano le stesse nei due gruppi. Sulla base delle assunzioni della proposizione 1, la componente Δ^{ν}_{X} riflette le differenze tra le distribuzioni delle caratteristiche osservate (X) dei due gruppi, mentre la componente della struttura salariale Δ^{ν}_{S} riflette le differenze di remunerazione rispetto sia a X che a ε.

La scomposizione aggregata della proposizione 1 è applicabile alla differenza di qualsiasi statistica distribuzionale, però è valida unicamente nel

supporto comune. Qui di seguito presentiamo il caso particolare della differenza tra le medie salariali di due gruppi, che permette di scomporre la differenza osservata tenendo conto anche degli individui che si trovano fuori dal supporto comune. Questo tipo di scomposizione, presentata da Nopo (2008), introduce due ulteriori termini alla scomposizione:

$$\begin{aligned}\Delta_O^\mu &= \mu(F_{Y_A|D_A}) - \mu(F_{Y_B|D_B}) = \\ &= E(Y_A) - E(Y_B) \\ &= \underbrace{p_{\mathcal{S}^C|A} \cdot \left(E(Y_{A|\mathcal{S}^C}) - E(Y_{A|\mathcal{S}})\right)}_{\Delta_A^\mu} + \underbrace{\left(E(Y_{A|\mathcal{S}}) - E(Y_{B|\mathcal{S}})\right)}_{\Delta_{O|\mathcal{S}}^\mu = \Delta_X^\mu + \Delta_S^\mu} + \underbrace{p_{\mathcal{S}^C|B} \cdot \left(E(Y_{B|\mathcal{S}}) - E(Y_{B|\mathcal{S}^C})\right)}_{\Delta_B^\mu},\end{aligned} \tag{2.9}$$

dove:

- $p_{\mathcal{S}^C|A} = Pr(X \in \mathcal{S}^C | D_A = 1)$ è la probabilità che le caratteristiche osservate non fanno parte del supporto comune sapendo che gli individui appartengono al gruppo A;
- $E(Y_{A|\mathcal{S}^C})$ è la media dei salari degli individui del gruppo A che non fanno parte del supporto comune;
- $E(Y_{A|\mathcal{S}})$ è il salario medio degli individui del gruppo A che appartengono al supporto comune;
- $E(Y_{B|\mathcal{S}})$ è il salario medio degli individui del gruppo B che appartengono al supporto comune;
- $p_{\mathcal{S}^C|B} = Pr(X \in \mathcal{S}^C | D_B = 1)$ è la probabilità che le caratteristiche osservate non appartengono al supporto comune sapendo che gli individui fanno parte del gruppo B;
- il termine centrale, $\Delta_{O|\mathcal{S}}^\mu$, non è nient'altro che la differenza tra i salari medi dei gruppi A e B nel **supporto comune** e può essere scomposto in due componenti, $\Delta_X^\mu + \Delta_S^\mu$, come nella proposizione 1.

L'equazione scompone la differenza osservata tra le medie salariali di tutti gli individui dei gruppi A e B nelle quattro componenti:

$$\Delta_O^\mu = \Delta_A^\mu + \underbrace{\Delta_X^\mu + \Delta_S^\mu}_{\Delta_{O|S}^\mu} + \Delta_B^\mu. \tag{2.10}$$

In questa scomposizione appaiono i due nuovi termini Δ_A^μ e Δ_B^μ:

- Δ_A^μ è la parte della differenza osservata imputabile al fatto che ci sono degli individui del gruppo A fuori dal supporto comune che hanno una media salariale diversa rispetto ai membri del gruppo A con caratteristiche del supporto comune;
- Δ_B^μ è la parte della differenza osservata dovuta al fatto che ci sono degli individui del gruppo B fuori dal supporto comune che hanno una media salariale diversa rispetto ai membri del gruppo B con caratteristiche del supporto comune.

La scomposizione proposta da Nopo (2008) tiene conto esplicitamente della questione del supporto comune dei due gruppi e permette di scomporre la differenza osservata tra le medie salariali dei due gruppi. Dei quattro termini della scomposizione, Δ_S^μ rappresenta la parte non spiegata, non giustificabile dalle caratteristiche diverse che possono avere i due gruppi. Gli altri tre termini, invece, costituiscono la parte spiegata:

- Δ_A^μ e Δ_B^μ sono le parti della differenza attribuibili al fatto che ci possono essere persone di un gruppo con delle caratteristiche che nessuno dell'altro gruppo ha;

- Δ_X^μ è la parte di differenza che si può imputare alla diversa distribuzione delle caratteristiche dei due gruppi, nell'insieme delle combinazioni di caratteristiche in cui si osservano individui di entrambi i gruppi (il supporto comune).

La scomposizione della differenza tra le medie salariali può quindi anche essere riscritta in questo modo:

$$\Delta_O^\mu = (\Delta_A^\mu + \Delta_B^\mu + \Delta_X^\mu) + \Delta_S^\mu.$$

Con questa formulazione, la differenza osservata tra le medie salariali può essere interpretata come si faceva tradizionalmente con la scomposizione di BO: una parte spiegabile dalle diverse caratteristiche osservate, e una parte non spiegata.

Nelle prossime sezioni presenteremo in un primo momento un metodo empirico per la scomposizione delle differenze salariali, e in un secondo momento degli esempi con dei dati fittizi.

3.
PARTE EMPIRICA

3.1 Introduzione

Nelle sezioni precedenti sono state formalizzate la scomposizione aggregata delle differenze salariali tra due gruppi e le sue condizioni di identificazione. Questo è il quadro teorico che costituisce il punto di partenza per varie metodologie empiriche per effettuare le scomposizioni. In questa sezione ne presentiamo una, basata sulla rielaborazione della funzione di ripartizione dei salari proposta da DiNardo, Fortin e Lemieux (1996) e, per quanto riguarda la riponderazione, sul metodo CEM (coarsened exact matching) introdotto da Iacus, King e Porro (2011). Una particolarità del metodo che proponiamo è che sarà applicabile anche a inchieste con piani di campionamento complessi. Per il caso particolare della media, presentiamo il metodo con la scomposizione proposta da Nopo (2008).

Come detto nelle sezioni precedenti, l'obiettivo è quello di scomporre le differenze salariali osservate tra due gruppi, in termini di una statistica distribuzionale, in una parte spiegata (dalle differenze nelle distribuzioni delle caratteristiche osservate, X) e in una parte non spiegata (attribuibile a una differenza nelle strutture salariali dei due gruppi).

Indichiamo con $\mathbf{X} = (X_1, X_2, \ldots, X_k)$ un insieme di dati k-dimensionale, dove ogni X_j è un vettore colonna di caratteristiche osservate per gli n individui del campione. Più precisamente, si può esprimere l'insieme k-dimensionale dei dati, $\mathbf{X}$, come $\mathbf{X} = [X_{ij}, i = 1, \ldots, n; j = 1, \ldots, k]$. Per ogni individuo i del campione si osserva quindi un vettore (riga) di k caratteristiche, $\mathbf{X}_i = (X_{i1}, \ldots, X_{ik})$.

Per ogni individuo del campione, oltre a delle caratteristiche, si osserva un salario, Y_i, e un individuo appartiene a uno dei due gruppi $g = A, B$. Riprendendo la notazione utilizzata nell'introduzione (gruppi mutualmente esclusivi), il salario dell'individuo i, Y_i, può essere espresso come $Y_i = D_{Ai}Y_{Ai}$, se l'individuo fa parte del gruppo A, o come $Y_i = D_{Bi}Y_{Bi}$ se appartiene al gruppo B.

Ogni osservazione possiede inoltre un peso, W_i, che deriva dal piano di campionamento.

Il punto cruciale per poter effettuare una scomposizione consiste nel riuscire a stimare una distribuzione controfattuale come quella dell'equazione (2.6), che riproponiamo qui con una notazione alleggerita:

$$F_{Y_A^C|\mathcal{S}}(y) = \int F_{Y_A|X_A,\mathcal{S}}(y|X) \cdot dF_{X_B|\mathcal{S}}(X). \tag{3.1}$$

Ci sono vari approcci per stimare la distribuzione controfattuale $F_{Y_A^C|\mathcal{S}}(y)$[1]. Quello che proponiamo in questo documento si basa sulla rielaborazione dell'equazione (3.1) proposta da DiNardo, Fortin e Lemieux (1996) (che chiameremo DFL da qui in avanti):

$$F_{Y_A^C|\mathcal{S}}(y) = \int F_{Y_A|X_A,\mathcal{S}}(y|X) \cdot \Psi_A(X) \cdot dF_{X_A|\mathcal{S}}(X), \tag{3.2}$$

dove $\Psi_A(X) = dF_{X_B|\mathcal{S}}(X)/dF_{X_A|\mathcal{S}}(X)$ è un fattore di riponderazione. Questa rielaborazione rende chiaro il fatto che la distribuzione controfattuale $F_{Y_A^C|\mathcal{S}}(\cdot)$ è semplicemente una versione riponderata della distribuzione dei salari del gruppo A, $F_{Y_A|\mathcal{S}}(\cdot)$. Il fattore di riponderazione, $\Psi_A(X)$, è una funzione che dipende unicamente dalla distribuzione delle caratteristiche osservate nei due gruppi, A e B, e una sua stima, come vedremo in

[1] In alternativa, si può utilizzare anche l'altra distribuzione controfattuale, ossia, $F_{Y_B^C|\mathcal{S}}(y) = \int F_{Y_B|X_B,\mathcal{S}}(y|X) \cdot dF_{X_A|\mathcal{S}}(X)$. Questa mescola la struttura dei salari del gruppo B con la distribuzione delle caratteristiche del gruppo A.

seguito, è una questione che non pone particolari problemi. L'idea di base dell'approccio DFL è di iniziare con il gruppo A, e in seguito sostituire la distribuzione delle caratteristiche osservate X del gruppo A ($F_{X_A|\mathcal{S}}(\cdot)$) con la distribuzione di X del gruppo B ($F_{X_B|\mathcal{S}}(\cdot)$), utilizzando il fattore di riponderazione $\Psi_A(X)$.

La distribuzione controfattuale dell'equazione (3.1), che rappresenta la distribuzione dei salari che avrebbero i lavoratori del gruppo A se le loro caratteristiche fossero distribuite come quelle del gruppo B, viene costruita, come mostrato nell'equazione (3.2), utilizzando il fattore di riponderazione:

$$\Psi_A(X) = \frac{dF_{X_B|\mathcal{S}}(X)}{dF_{X_A|\mathcal{S}}(X)}. \tag{3.3}$$

Il fattore di riponderazione è il rapporto tra le distribuzioni congiunte delle caratteristiche dei due gruppi nel supporto comune. Considerando unicamente variabili discrete, le distribuzioni congiunte possono essere stimate semplicemente con le frequenze relative congiunte delle caratteristiche dei due gruppi nel supporto comune.

Indicando con $\widehat{f}_{X_A|\mathcal{S}}$ le frequenze relative congiunte delle caratteristiche del gruppo A e rispettivamente $\widehat{f}_{X_B|\mathcal{S}}$ per il gruppo B, il fattore di riponderazione da applicare all'individuo i del gruppo A può essere stimato nella maniera seguente:

$$\widehat{\Psi}_A(\mathbf{x}_i) = \frac{\widehat{f}_{X_B|\mathcal{S}}(\mathbf{x}_i)}{\widehat{f}_{X_A|\mathcal{S}}(\mathbf{x}_i)}, \tag{3.4}$$

dove $\mathbf{x}_i = (x_{i1}, \ldots, x_{ik})$ è il vettore delle k caratteristiche osservate dell'individuo i e $\widehat{f}_{X_A|\mathcal{S}}$ e $\widehat{f}_{X_B|\mathcal{S}}$ sono le frequenze relative congiunte delle caratteristiche dei gruppi A e B nel supporto comune, stimate in questo modo:

$$\widehat{f}_{X_g|\mathcal{S}}(\mathbf{x}_i) = \frac{\sum\limits_{\substack{j \in g \\ X_j \in \mathcal{S}}} W_j \cdot \mathbf{1}(X_{j1} = x_{i1}, \ldots, X_{jk} = x_{ik})}{\sum\limits_{\substack{j \in g \\ X_j \in \mathcal{S}}} W_j},$$

dove W_j sono i pesi di campionamento e $g = A, B$.

Una volta stimati i fattori di riponderazione per ogni individuo del gruppo A nel supporto comune, si hanno a disposizione tutti gli elementi per poter effettuare la scomposizione delle differenze tra le statistiche salariali di due gruppi di individui osservate nel supporto comune.

In pratica i fattori di riponderazione vengono utilizzati per stimare la statistica dei salari controfattuale, che è la quantità chiave per poter scomporre le differenze salariali nelle due componenti dell'equazione (2.8).

Nella prossima sezione presentiamo un esempio empirico su un campione di dati fittizi, in cui verranno mostrati tutti i passaggi dell'analisi che portano alla scomposizione della differenza di una statistica salariale tra due gruppi.

3.2 Esempio empirico

In questa sezione utilizziamo un campione fittizio di dati per analizzare le differenze salariali tra uomini e donne. Questi dati sono inclusi nel pacchetto R "decr" e le istruzioni per reperirli si trovano nell'appendice A. Il campione è composto da 1.000 individui, di cui 547 uomini e 453 donne. Per ogni osservazione si hanno a disposizione le informazioni seguenti:

- Sesso: uomo o donna,
- Settore economico: settore economico in cui l'individuo lavora (secondario o terziario),
- Grado di formazione: grado di formazione della persona (I, II o III),
- Salario mensile: salario mensile lordo in franchi, standardizzato a un tempo di lavoro a tempo pieno,
- Peso di campionamento: peso di campionamento associato a ogni osservazione. In pratica questo peso corrisponde al numero di osservazioni che l'individuo rappresenta della popolazione di riferimento. Questo peso viene considerato nella stima delle statistiche campionarie (media, mediana, numerosità, ecc...).

L'obiettivo di questa sezione è quello di analizzare le differenze che si osservano tra le statistiche salariali degli uomini e delle donne, applicando la metodologia presentata nella sezione 3.1 per ottenere le scomposizioni dell'equazione (2.10) per quanto riguarda le medie e dell'equazione (2.8) per i quantili.

Le prime righe del campione fittizio si presentano come nella tabella [T.3.1].

T. 3.1
Prime quattro righe del campione fittizio

Sesso	Settore economico	Grado formazione	Salario mensile	Peso
Uomini	Terziario	III	8.400	105
Donne	Secondario	II	4.200	32
Uomini	Terziario	II	5.100	36
Donne	Terziario	II	7.400	12

Fonte: Campione di dati fittizi; elaborazione Ustat

Ogni riga di questi dati rappresenta un salariato o una salariata inventati. La prima riga, per esempio, rappresenta un uomo che lavora nel settore terziario, con una formazione di grado terziario (III) e un salario mensile di 8.400 franchi.

Il salario medio stimato dal campione è di 4.645 franchi al mese. Distinguendo le persone secondo il genere, la media dei salari mensili degli uomini è di 5.323 franchi, mentre quella delle donne è di 3.614 franchi; si osserva quindi una differenza di 1.709 franchi a favore degli uomini [T. 3.2].

T. 3.2
Numerosità (n), stima degli effettivi ($\widehat{N}$) e salari medi ($\widehat{\overline{Y}}$, in franchi mensili), secondo il genere (totale del campione fittizio)

Sesso	n	$\widehat{N}$	$\widehat{\overline{Y}}$
Totale	1.000	70.196	4.645
Uomini	547	42.363	5.323
Donne	453	27.833	3.614
Differenze salariali osservate			**1.709**

Fonte: Campione di dati fittizi; elaborazione Ustat

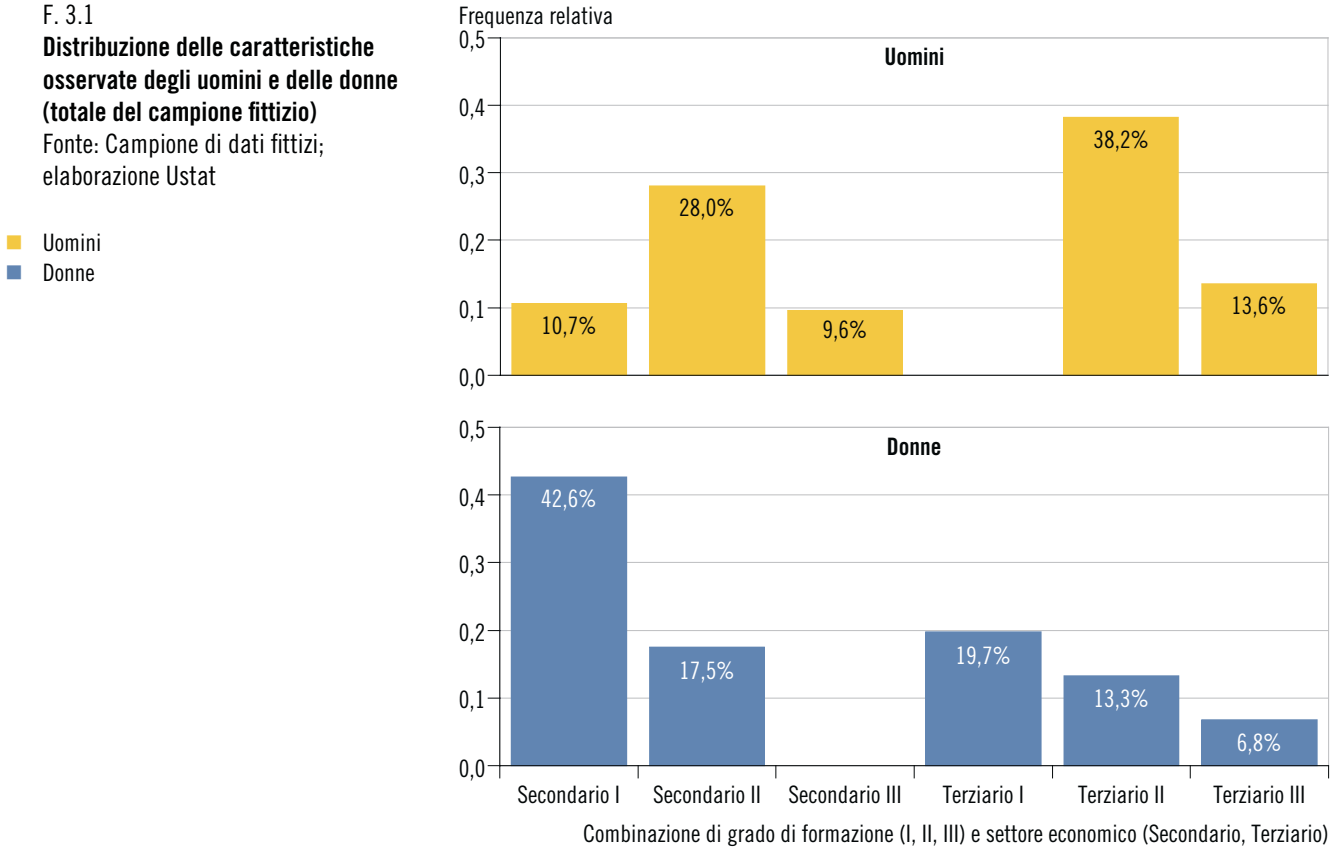
Questa è la differenza osservata tra il salario medio di tutti gli uomini del campione e il salario medio di tutte le donne del campione, senza tenere conto di eventuali differenze di caratteristiche tra i due gruppi. In questo esempio, per semplicità, prendiamo in considerazione solo due caratteristiche: il settore economico in cui la persona lavora e il suo grado di formazione. Come si distribuiscono gli uomini e le donne rispetto alle combinazioni di settore economico e grado di formazione? La risposta a questa domanda si ottiene osservando le frequenze relative congiunte delle caratteristiche degli uomini e delle donne [F. 3.1].

Il settore economico e il grado di formazione generano congiuntamente sei possibili combinazioni o strati. Dalla figura [F. 3.1] emergono due informazioni importanti:

- gli uomini e le donne non sono rappresentati negli stessi strati. Non ci sono donne con una formazione terziaria che lavorano nel secondario e non ci sono uomini con una formazione primaria che lavorano nel settore terziario;
- negli strati in cui si osservano sia uomini che donne, i due gruppi si distribuiscono in maniera diversa.

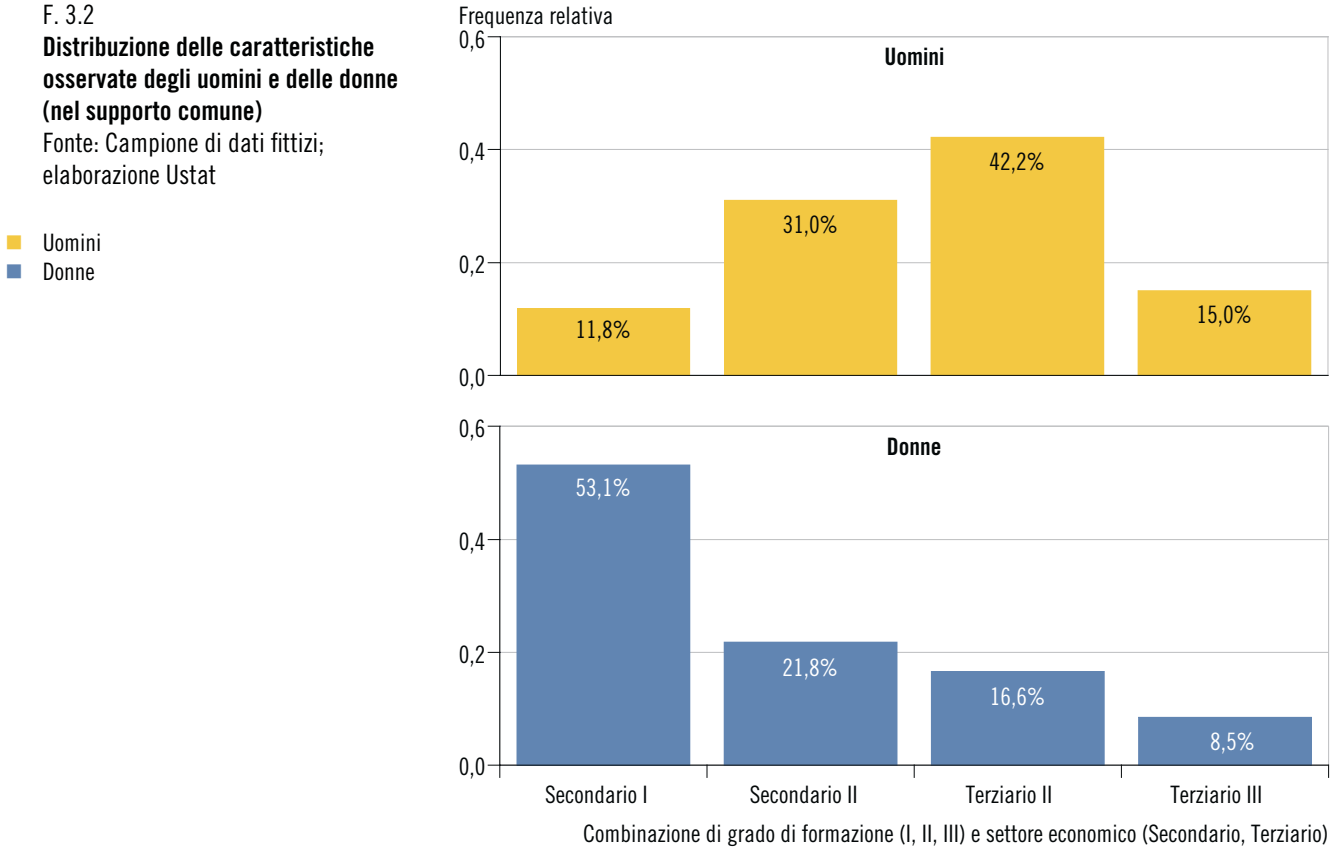
Il metodo che vogliamo applicare per analizzare la differenza osservata tra i salari medi tiene conto di questi due punti. In primo luogo, si stabilisce il supporto comune, che è definito dai quattro strati in cui si osservano sia uomini che donne. In secondo luogo, nella zona del supporto comune, si andranno a stimare i fattori di riponderazione dell'equazione (3.4), che permetteranno di bilanciare la distribuzione delle caratteristiche degli uomini a quella delle donne e di stimare un salario medio **controfattuale**, che è la quantità chiave per poter effettuare la scomposizione della differenza salariale tra le medie dei due gruppi.

F. 3.1
Distribuzione delle caratteristiche osservate degli uomini e delle donne (totale del campione fittizio)
Fonte: Campione di dati fittizi; elaborazione Ustat



Quindi ora focalizziamo l'attenzione sulla distribuzione congiunta delle caratteristiche di uomini e donne nei quattro strati che formano il **supporto comune** [F. 3.2]. Il supporto comune ha escluso lo strato "settore secondario - formazione III", in cui si osservano solo uomini, e lo strato "settore terziario - formazione I", in cui ci sono solo donne. Questi due strati sono stati esclusi dal supporto comune perché gli uomini e le donne che ne fanno parte non sono paragonabili. Ad esempio le donne con una formazione primaria e attive nel terziario non hanno nemmeno un uomo con le medesime caratteristiche a cui essere paragonate.

F. 3.2
Distribuzione delle caratteristiche osservate degli uomini e delle donne (nel supporto comune)
Fonte: Campione di dati fittizi; elaborazione Ustat



T. 3.3
Stima dei quattro fattori di riponderazione ($\widehat{\Psi}_A(X)$), per ogni strato del supporto comune

Strato	$\widehat{\Psi}_A(X)$
Secondario I	4,50
Secondario II	0,70
Terziario II	0,39
Terziario III	0,56

Fonte: Campione di dati fittizi; elaborazione Ustat

Nella figura [F. 3.2] ogni gruppo di donne trova tra gli uomini un gruppo con le stesse caratteristiche. Ora abbiamo uomini e donne negli stessi strati, ma con pesi diversi. Per rendere i due gruppi completamente paragonabili, introduciamo il fattore di riponderazione (vedi equazione (3.4)), che permette proprio di bilanciare la distribuzione congiunta delle caratteristiche degli uomini a quella delle donne. La stima di questo fattore di riponderazione non è nient'altro che, per ogni strato del supporto comune, il rapporto tra il peso delle donne e quello degli uomini della figura [F. 3.2].

Per esempio, il fattore di riponderazione per lo strato "settore secondario - formazione I" è uguale a $\frac{53{,}1\%}{11{,}8\%} = 4{,}5$. In questo esempio otteniamo così quattro fattori di riponderazione, che vanno poi associati a ogni uomo del supporto comune, in funzione dello strato a cui appartiene. I 4 fattori di riponderazione, uno per ogni strato del supporto comune, sono riportati nella tabella [T. 3.3].

Una volta associati i fattori di riponderazione agli uomini del supporto comune, siamo in grado di stimare qualsiasi statistica dei salari controfattuale. In questo primo esempio stimiamo il salario medio controfattuale, e cioè il salario medio che gli uomini del supporto comune avrebbero se avessero la stessa distribuzione delle caratteristiche delle donne, con la formula seguente:

$$\widehat{\overline{Y}}^{C}_{A|\mathcal{S}} = \frac{\sum\limits_{\substack{i\in A \\ X_i\in\mathcal{S}}} W_i \cdot \widehat{\Psi}_A(X_i) \cdot Y_i}{\sum\limits_{\substack{i\in A \\ X_i\in\mathcal{S}}} W_i \cdot \widehat{\Psi}_A(X_i)}, \tag{3.5}$$

dove W_i è il peso di campionamento dell'uomo i, Y_i il suo salario, e $\widehat{\Psi}_A(X_i)$ il fattore di riponderazione che gli è stato assegnato. Questa non è nient'altro che la stima di un salario medio ponderato, con l'inclusione del fattore di riponderazione che fa in modo che il risultato indichi una media salariale degli uomini "a parità di caratteristiche" con quella delle donne. Applicando la formula (3.5) agli uomini del supporto comune del nostro campione di dati fittizi, otteniamo un salario medio controfattuale di 5.006 franchi mensili. A questo punto siamo quasi pronti a svolgere la scomposizione della differenza osservata tra le medie salariali degli uomini e delle donne dei dati fittizi di questo esempio. A tal fine, servono ancora alcune stime, e in particolare:

- il salario medio degli uomini che non fanno parte del supporto comune e la stima della loro numerosità,
- il salario medio degli uomini del supporto comune e la stima della loro numerosità,
- il salario medio delle donne del supporto comune e la stima della loro numerosità,
- il salario medio delle donne fuori dal supporto comune e la stima della loro numerosità.

Queste stime, insieme al salario medio controfattuale, ci permettono di scomporre la differenza tra le medie salariali degli uomini e delle donne come nelle equazioni (2.9) e (2.10). Le stime delle quantità necessarie alla scomposizione sono presentate nella tabella [T. 3.4].

T. 3.4
Stime dei salari medi ($\widehat{\overline{Y}}$, in franchi mensili) e della numerosità ($\widehat{N}$) degli uomini e delle donne dentro e fuori dal supporto comune

Sesso	Supporto comune	$\widehat{\overline{Y}}$	$\widehat{N}$
Uomini	FALSO	6.057	4.056
Uomini	VERO	5.245	38.307
Donne	FALSO	3.369	5.496
Donne	VERO	3.674	22.337

Fonte: Campione di dati fittizi; elaborazione Ustat

E ora la scomposizione alla Nopo (2008) della differenza tra i salari medi di uomini e donne dell'equazione (2.9), riproposta qui sotto in forma di stimatori empirici:

$$\begin{aligned}\widehat{\Delta}^{\mu}_{O} &= \widehat{\overline{Y}}_A - \widehat{\overline{Y}}_B = \\ &= \underbrace{\widehat{p}_{\mathcal{S}^C|A} \cdot \left(\widehat{\overline{Y}}_{A|\mathcal{S}^C} - \widehat{\overline{Y}}_{A|\mathcal{S}}\right)}_{\widehat{\Delta}^{\mu}_{A}} + \underbrace{\left(\widehat{\overline{Y}}_{A|\mathcal{S}} - \widehat{\overline{Y}}_{B|\mathcal{S}}\right)}_{\widehat{\Delta}^{\mu}_{O|\mathcal{S}}=\widehat{\Delta}^{\mu}_{X}+\widehat{\Delta}^{\mu}_{S}} + \underbrace{\widehat{p}_{\mathcal{S}^C|B} \cdot \left(\widehat{\overline{Y}}_{B|\mathcal{S}} - \widehat{\overline{Y}}_{B|\mathcal{S}^C}\right)}_{\widehat{\Delta}^{\mu}_{B}},\end{aligned} \tag{3.6}$$

dove:

- $\widehat{\overline{Y}}_A$ e $\widehat{\overline{Y}}_B$ sono i salari medi osservati di uomini e donne, rispettivamente (5.323 e 3.614 franchi mensili nel nostro esempio, vedi tabella [T. 3.2]),
- $\widehat{p}_{\mathcal{S}^C|A}$ è la proporzione di uomini fuori dal supporto comune (rispetto a tutti gli uomini del campione). Nel nostro esempio equivale a $\frac{4.056}{4.056+38.307} = 9,58\%$ (cifre riprese dalla tabella [T. 3.4]),
- $\widehat{\overline{Y}}_{A|\mathcal{S}^C}$ è il salario medio degli uomini che non fanno parte del supporto comune (6.057 franchi mensili nel nostro esempio, vedi tabella [T. 3.4]),
- $\widehat{\overline{Y}}_{A|\mathcal{S}}$ è il salario medio degli uomini che fanno parte del supporto comune (5.245 franchi mensili nel nostro esempio, vedi tabella [T. 3.4]),
- $\widehat{\overline{Y}}_{B|\mathcal{S}}$ è il salario medio delle donne che fanno parte del supporto comune (3.674 franchi mensili nel nostro esempio, vedi tabella [T. 3.4]),
- $\widehat{\overline{Y}}_{B|\mathcal{S}^C}$ è il salario medio delle donne che non fanno parte del supporto comune (3.369 franchi mensili nel nostro esempio, vedi tabella [T. 3.4]),
- $\widehat{p}_{\mathcal{S}^C|B}$ è la proporzione di donne fuori dal supporto comune (rispetto a tutte le donne del campione). Nel nostro esempio equivale a $\frac{5.496}{5.496+22.337} = 19,75\%$ (cifre riprese dalla tabella [T. 3.4]).

Il salario medio controfattuale che abbiamo stimato poco sopra (5.006 franchi mensili) viene sottratto e aggiunto alla parte centrale dell'equazione (3.6), per scomporre la differenza salariale nel supporto comune in due componenti:

$$\widehat{\Delta}^{\mu}_{O|\mathcal{S}} = \underbrace{\left(\widehat{\overline{Y}}_{A|\mathcal{S}} - \widehat{\overline{Y}}^{C}_{A|\mathcal{S}}\right)}_{\widehat{\Delta}^{\mu}_{X}\text{ (spiegata)}} + \underbrace{\left(\widehat{\overline{Y}}^{C}_{A|\mathcal{S}} - \widehat{\overline{Y}}_{B|\mathcal{S}}\right)}_{\widehat{\Delta}^{\mu}_{S}\text{ (non spiegata)}}, \tag{3.7}$$

dove:

- $\widehat{\Delta}^{\mu}_{X}$ è la parte di differenza nel supporto comune giustificabile dal fatto che gli uomini e le donne hanno una distribuzione delle caratteristiche diversa (parte "spiegata"),
- $\widehat{\Delta}^{\mu}_{S}$ è la parte di differenza nel supporto comune non spiegabile dalla diversa distribuzione delle caratteristiche degli uomini e delle donne (parte "non spiegata").

La differenza osservata (totale) tra le medie salariali di uomini e donne viene così scomposta nelle quattro componenti:

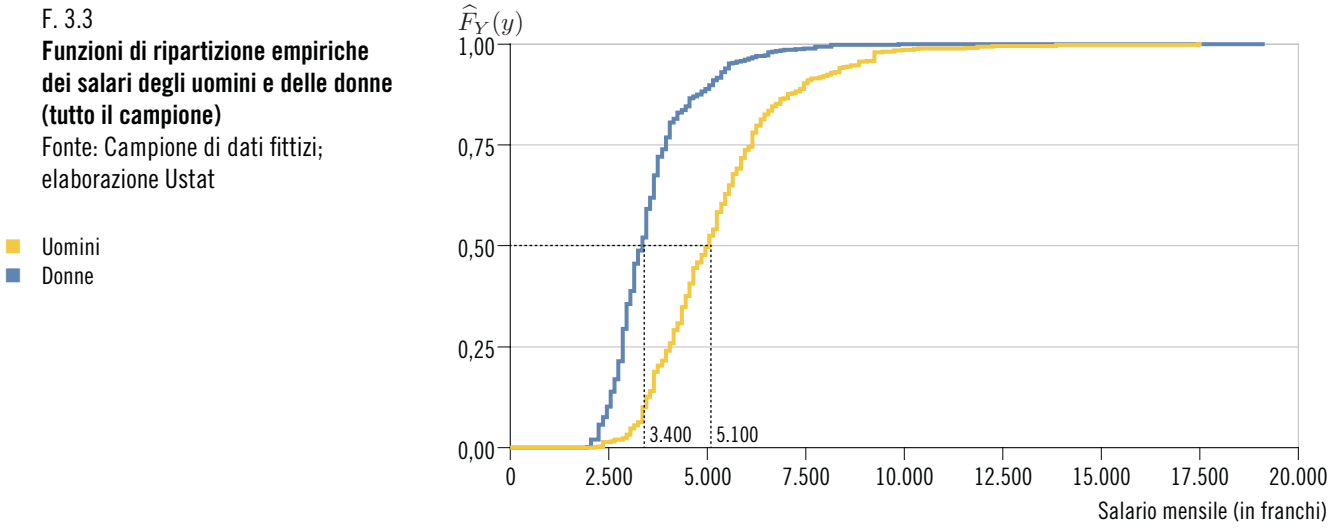
$$\begin{aligned}1.709 &= (5.323 - 3.614)\,\text{fr./mese} \\ &= 9,58\% \cdot (6.057 - 5.245) + (5.245 - 5.006) + (5.006 - 3.674) + 19,75\% \cdot (3.674 - 3.369) \\ &= \underbrace{77,7}_{\widehat{\Delta}^{\mu}_{A}} + \underbrace{239}_{\widehat{\Delta}^{\mu}_{X}} + \underbrace{1.332}_{\widehat{\Delta}^{\mu}_{S}} + \underbrace{60,3}_{\widehat{\Delta}^{\mu}_{B}}.\end{aligned}$$

La differenza osservata tra i salari medi di uomini e donne, come già detto, è di 1.709 franchi al mese ($\widehat{\Delta}_O^{\mu}$). Se teniamo conto congiuntamente delle caratteristiche formazione e settore economico di uomini e donne, rimane una differenza non spiegabile di 1.332 franchi mensili ($\widehat{\Delta}_S^{\mu}$). Dalla differenza osservata totale rimangono 1.709 – 1.332 = 377 franchi che si possono spiegare dal fatto che gli uomini e le donne hanno una distribuzione congiunta delle caratteristiche considerate differente. Questi 377 franchi di parte spiegata sono a loro volta scomposti in tre componenti che si addizionano tra loro:

- 77,7 franchi sono dovuti al fatto che ci sono degli uomini con delle caratteristiche che nessuna donna ha e che guadagnano in media di più degli uomini per i quali esistono delle donne con caratteristiche paragonabili ($\widehat{\Delta}_A^{\mu}$),
- 239 franchi sono dovuti al fatto che nei segmenti di caratteristiche nei quali esistono uomini e donne con caratteristiche paragonabili (supporto comune) la distribuzione congiunta delle caratteristiche dei due gruppi è diversa ($\widehat{\Delta}_X^{\mu}$),
- 60,3 franchi sono invece spiegabili dal fatto che esiste un gruppo di donne con delle caratteristiche per le quali non esiste nessun uomo e che guadagnano mediamente di meno rispetto alle donne del supporto comune ($\widehat{\Delta}_B^{\mu}$).

Ora che abbiamo scomposto la differenza osservata tra le medie salariali di uomini e donne, possiamo approfondire l'analisi con le scomposizioni delle differenze tra i quantili salariali dei due gruppi, come nell'equazione (2.8). Ricordiamo che queste scomposizioni sono possibili solo nel supporto comune. L'analisi mira in particolare a come si distribuisce la parte non spiegata delle differenze salariali, per verificare se è costante oppure se ci sono delle differenze tra la parte bassa e quella alta della distribuzione. Prima di procedere con le scomposizioni delle differenze osservate tra i quantili dei salari nel supporto comune, diamo un'occhiata alle funzioni di ripartizione empiriche dei salari degli uomini e delle donne [F. 3.3].

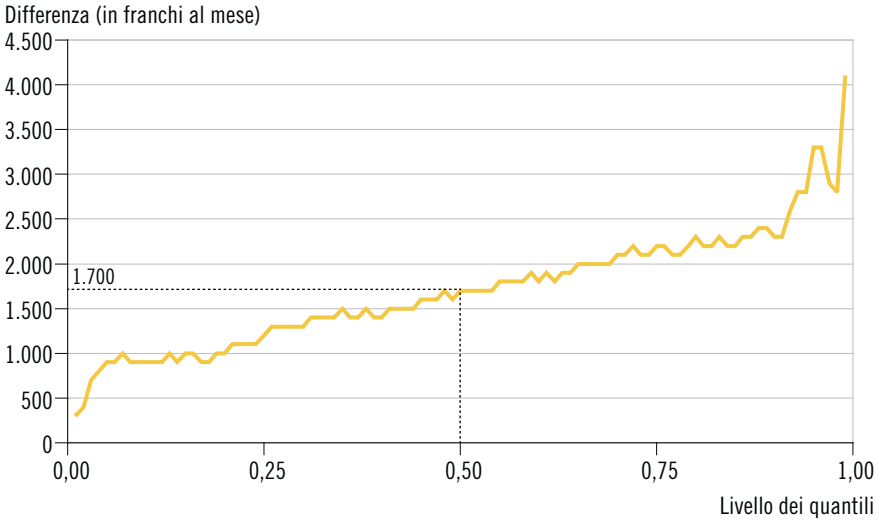
F. 3.3
Funzioni di ripartizione empiriche dei salari degli uomini e delle donne (tutto il campione)
Fonte: Campione di dati fittizi; elaborazione Ustat



Per ogni livello di salario rappresentato sull'asse delle ascisse della figura [F. 3.3], l'asse delle ordinate riporta la proporzione di uomini (linea gialla) e di donne (linea blu) con un salario uguale o inferiore a quel salario. Per esempio, dal grafico emerge che un uomo su due guadagna fino a 5.100 franchi al mese, mentre una donna su due guadagna 3.400 franchi o meno.

I due punti evidenziati nella figura [F. 3.3] rappresentano i salari mediani dei due gruppi. I quantili sono infatti la funzione inversa della funzione di ripartizione dei salari (o funzione di distribuzione cumulata). Se si considerano i salari mediani, la differenza che si osserva tra i due gruppi è quindi uguale a 5.100 – 3.400 = 1.700. Se ci focalizziamo sulle differenze tra i quantili dei salari a vari livelli, oltre che alla mediana, si ottiene la figura [F. 3.4], che riporta sull'asse delle ordinate le differenze osservate tra i percentili salariali degli uomini e delle donne.

F. 3.4
Differenza osservata tra i quantili dei salari degli uomini e delle donne (tutto il campione)
Fonte: Campione di dati fittizi; elaborazione Ustat

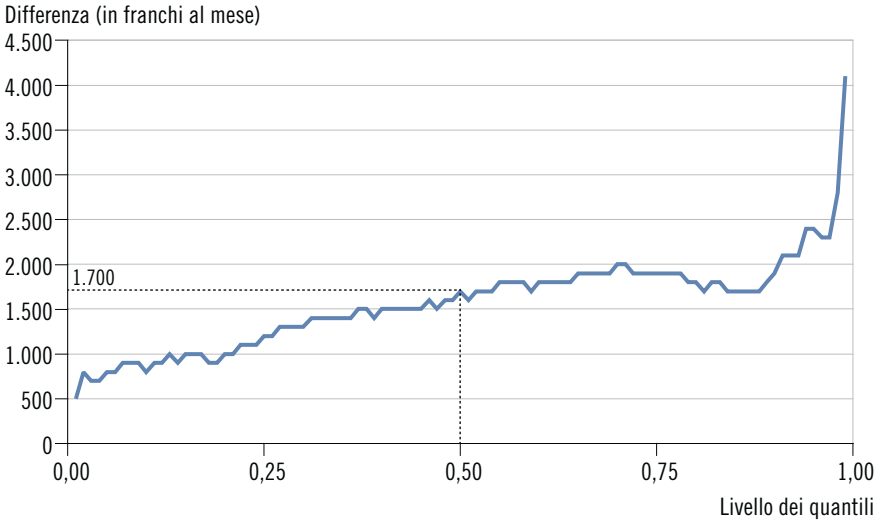


Il punto evidenziato nella figura [F. 3.4] riporta "verticalmente" la differenza tra i due punti evidenziati nella figura [F. 3.3] che era rappresentata "orizzontalmente".

La linea gialla della figura [F. 3.4] rappresenta le differenze osservate tra i quantili dei salari degli uomini e delle donne considerando tutto il campione fittizio. L'asse delle x riporta i livelli dei quantili (dal primo al novantanovesimo percentile), mentre l'asse delle y raffigura, per ogni percentile, la differenza tra i salari degli uomini e delle donne. Le linee punteggiate nel grafico evidenziano la differenza osservata tra le mediane salariali dei due gruppi (1.700 franchi al mese). Una prima osservazione che si può fare è che le differenze salariali tra uomini e donne non sono costanti e seguono un percorso piuttosto crescente al crescere del livello della distribuzione. Le differenze partono da 300 franchi al primo percentile (punto più a sinistra della figura [F. 3.4]) e raggiungono i 4.100 franchi al mese al novantanovesimo percentile (punto più a destra della figura [F. 3.4]), passando da una differenza a livello mediano di 1.700 franchi al mese.

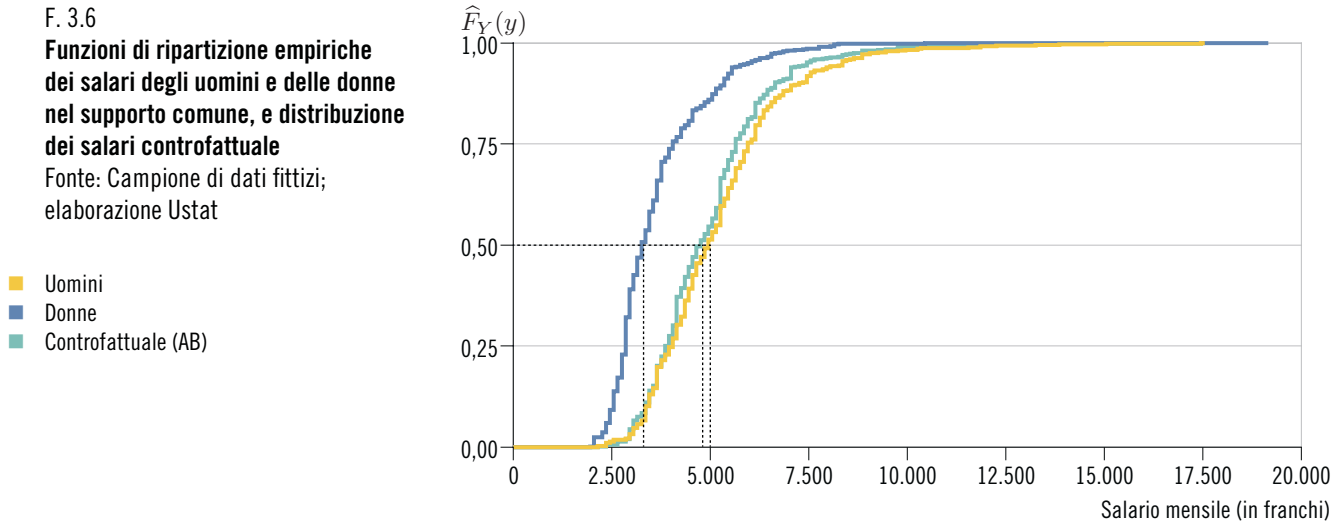
Queste differenze non possono essere scomposte come nell'equazione (2.8); la scomposizione può essere fatta solo all'interno del supporto comune. Se si considera solo il supporto comune, le differenze tra i quantili dei salari sono quelle rappresentate nella figura [F. 3.5].

F. 3.5
Differenza osservata tra i quantili dei salari degli uomini e delle donne (nel supporto comune)
Fonte: Campione di dati fittizi; elaborazione Ustat



A ogni punto della figura [F. 3.5] possiamo applicare la scomposizione dell'equazione (2.8), utilizzando gli stessi fattori di riponderazione usati per la scomposizione della differenza tra le medie salariali dell'esempio precedente. Con i fattori di riponderazione si può stimare una distribuzione dei salari controfattuale, e cioè una distribuzione dei salari degli uomini, nel supporto comune, con la stessa distribuzione delle caratteristiche delle donne. La figura [F. 3.6] presenta le distribuzioni cumulate empiriche dei salari degli uomini e delle donne nel supporto comune e la distribuzione dei salari controfattuale.

F. 3.6
Funzioni di ripartizione empiriche dei salari degli uomini e delle donne nel supporto comune, e distribuzione dei salari controfattuale
Fonte: Campione di dati fittizi; elaborazione Ustat



Le tre funzioni di ripartizione dei salari della figura [F. 3.6] contengono tutte le informazioni necessarie per effettuare qualsiasi scomposizione delle differenze salariali tra le statistiche distribuzionali di due gruppi, come nell'equazione (2.8).

Se si prende un livello di quantile sull'asse verticale della figura [F. 3.6], la scomposizione della differenza tra i quantili dei salari degli uomini e delle donne è data da:

- la differenza "orizzontale" tra la linea gialla e la linea verde, che rappresenta la parte spiegata della differenza osservata nel supporto comune e
- la differenza "orizzontale" tra la linea verde e la linea blu, che rappresenta la parte non spiegata della differenza osservata nel supporto comune.

Nella figura [F. 3.6] sono evidenziati i punti per la scomposizione della differenza dei salari mediani di uomini e donne nel supporto comune:

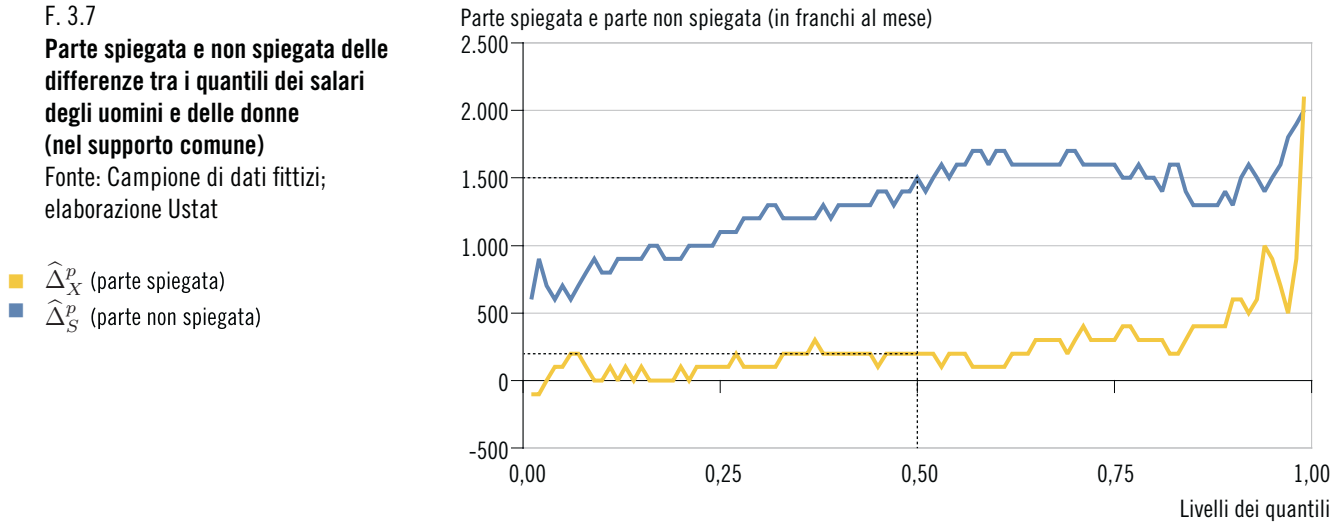
$$5.000 - 3.300 = \underbrace{(5.000 - 4.800)}_{\text{parte spiegata } (\widehat{\Delta}_X^{p50})} + \underbrace{(4.800 - 3.300)}_{\text{parte non spiegata } (\widehat{\Delta}_S^{p50})} = \underbrace{200}_{\widehat{\Delta}_X^{p50}} + \underbrace{1.500}_{\widehat{\Delta}_S^{p50}}.$$

In pratica, la differenza di 1.700 franchi che si osserva tra il salario mensile mediano degli uomini e delle donne nel supporto comune viene scomposta in:

- 200 franchi spiegabili dal fatto che gli uomini e le donne hanno due distribuzioni diverse delle caratteristiche considerate (settore economico e grado di formazione)
- 1.500 franchi non spiegabili (o giustificabili) dalle diverse distribuzioni delle caratteristiche dei due gruppi (nel supporto comune).

Se ripetiamo la scomposizione della differenza tra i salari mediani a tutti i percentili, dal primo al novantanovesimo, otteniamo i risultati raffigurati nella figura [F. 3.7]. Le due linee della figura [F. 3.7] rappresentano la parte spiegata e non spiegata delle differenze osservate nel supporto comune tra i percentili dei salari degli uomini e delle donne, la cui somma "verticale" equivale alla linea della figura [F. 3.5].

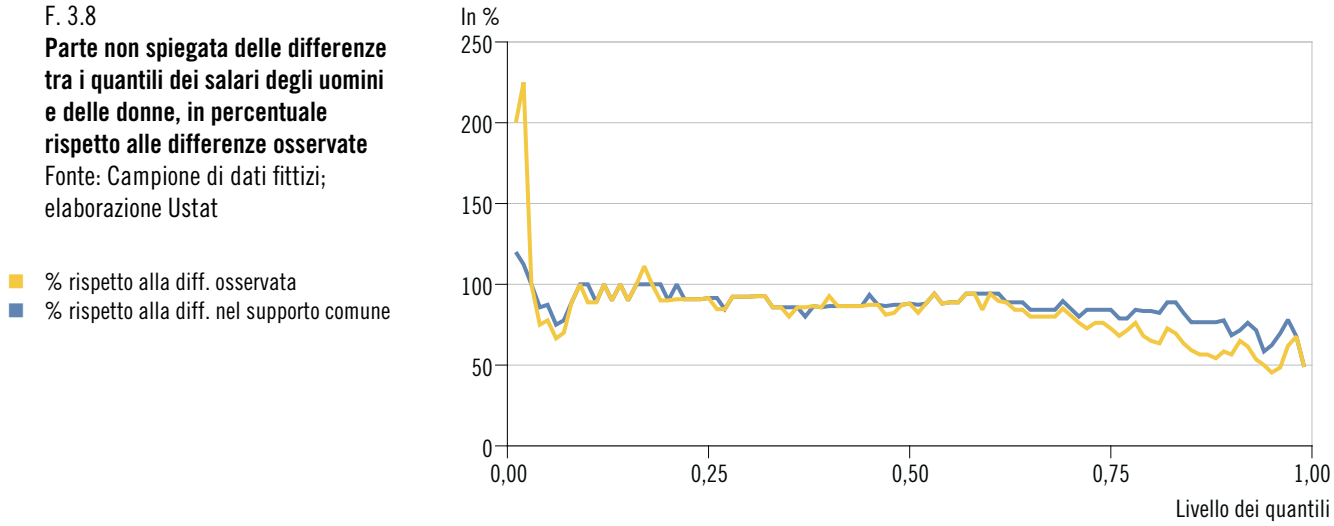
F. 3.7
Parte spiegata e non spiegata delle differenze tra i quantili dei salari degli uomini e delle donne (nel supporto comune)
Fonte: Campione di dati fittizi; elaborazione Ustat



La parte non spiegata delle differenze tra i percentili salariali degli uomini e delle donne (linea blu della figura [F. 3.7]), che rappresenta le differenze salariali "a parità di caratteristiche", rimane positiva (in favore degli uomini) lungo tutta la distribuzione dei salari. Le differenze non spiegate sono piuttosto crescenti dal primo fino a circa il sessantesimo percentile, dove raggiungono i 1.700 franchi al mese. Tra il sessantesimo e il novantacinquesimo percentile, la parte non spiegata delle differenze oscilla tra i 1.300 e i 1.500 franchi al mese, per poi tornare a crescere fino a 2.000 franchi mensili al novantanovesimo percentile. Questi risultati mostrano come le differenze salariali tra uomini e donne non sono costanti, anche dopo aver considerato le differenze tra le distribuzioni delle caratteristiche dei due gruppi. Inoltre, nella parte bassa della distribuzione, le differenze non spiegate non sono molto diverse dalle differenze osservate (sia considerando tutto il campione che quelle osservate solo nel supporto comune). Nella parte alta della distribuzione, invece, la parte non spiegata delle differen-

ze, pur rimanendo importante, si riduce rispetto alle differenze osservate. Questo può essere verificato considerando la parte non spiegata in percentuale rispetto alle differenze osservate, sia totali che solo nel supporto comune, come riportato nella figura [F. 3.8].

F. 3.8
Parte non spiegata delle differenze tra i quantili dei salari degli uomini e delle donne, in percentuale rispetto alle differenze osservate
Fonte: Campione di dati fittizi; elaborazione Ustat



Nei primi due percentili, la parte non spiegata supera addirittura il 100% della differenza osservata, diventando anche più del doppio. Fatta eccezione per i primi due percentili, fino a circa il sessantesimo percentile la parte non spiegata non si distanzia di molto rispetto alla differenza osservata, oscillando tra il 75% e il 110% circa della differenza complessiva osservata. Dopo il sessantesimo percentile, invece, la parte non spiegata scende fino a circa il 50% della differenza complessiva osservata, nella parte più alta della distribuzione.

Nell'appendice A si trova il codice R per poter replicare l'esempio empirico presentato in questa sezione, con il pacchetto "decr" che abbiamo sviluppato.

APPENDICE A

Codice R per replicare l'esempio empirico, con il pacchetto "decr"

Codice R per replicare l'esempio empirico della sezione 3.2, con il pacchetto "decr".

```
# Installazione del pacchetto decr
install.packages("devtools")
devtools::install_github("gibonet/decr", build_vignettes = TRUE)
```

```
library(decr)
salari_inventati <- invented_wages
str(salari_inventati)
```

```
## Classes 'tbl_df' and 'data.frame':   1000 obs. of  5 variables:
##  $ gender        : Factor w/ 2 levels "men","women": 1 2 1 2 1 1 1 2 2 2 ..
##  $ sector        : Factor w/ 2 levels "secondary","tertiary": 2 1 2 2 1 1..
##  $ education     : Factor w/ 3 levels "I","II","III": 3 2 2 2 2 1 3 1 2 2..
##  $ wage          : num  8400 4200 5100 7400 4300 4900 5400 2900 4500 3000..
##  $ sample_weights: num  105 32 36 12 21 46 79 113 34 32 ...
```

```
# Traduzione dei dati dall'inglese all'italiano
colnames(salari_inventati) <- c("sesso", "settore_economico", "grado_formazione",
                                "salario_mensile", "peso")
levels(salari_inventati$sesso) <- c("Uomini", "Donne")
levels(salari_inventati$settore_economico) <- c("Secondario", "Terziario")

# Struttura dei dati
str(salari_inventati)
```

```
## Classes 'tbl_df' and 'data.frame':   1000 obs. of  5 variables:
##  $ sesso            : Factor w/ 2 levels "Uomini","Donne": 1 2 1 2 1 1 1 ..
##  $ settore_economico: Factor w/ 2 levels "Secondario","Terziario": 2 1 2 ..
##  $ grado_formazione : Factor w/ 3 levels "I","II","III": 3 2 2 2 2 1 3 1 ..
##  $ salario_mensile  : num  8400 4200 5100 7400 4300 4900 5400 2900 4500 3..
##  $ peso             : num  105 32 36 12 21 46 79 113 34 32 ...
```

```
# Prime righe dei dati
head(salari_inventati)
```

```
## # A tibble: 6 x 5
##   sesso  settore_economico grado_formazione salario_mensile  peso
##   <fctr> <fctr>            <fctr>                     <dbl> <dbl>
## 1 Uomini Terziario         III                         8400 105
## 2 Donne  Secondario        II                          4200  32.0
## 3 Uomini Terziario         II                          5100  36.0
## 4 Donne  Terziario         II                          7400  12.0
## 5 Uomini Secondario        II                          4300  21.0
## 6 Uomini Secondario        I                           4900  46.0
```

```
# Ogni riga di questi dati rappresenta un salariato o una salariata inventati,
# di cui si conoscono le informazioni contenute nelle colonne:
# - sesso: il sesso della persona ("Uomini" o "Donne")
# - settore_economico: il settore economico in cui la persona lavora
#   ("Secondario"o "Terziario")
# - grado_formazione: il grado di formazione della persona
#   ("I" = primario, "II" = secondario, "III" = terziario)
# - salario_mensile: salario mensile del lavoratore o della lavoratrice
# - peso: il peso che deriva dal piano di campionamento.

# Al fine di effettuare  una scomposizione della differenza osservata tra le
# statistiche distribuzionali di una variabile numerica (y) di uomini e donne,
# si procede con i passaggi seguenti:
# - Si stabilisce il supporto comune dei due gruppi, in termini di alcune
#   caratteristiche. In questo esempio, si considerano il settore economico in
#   cui un individuo lavora (settore_economico) e il grado di formazione della
#   persona (grado_formazione). Ogni combinazione osservata delle
#   caratteristiche costituisce uno `strato`, in cui si possono osservare:
#     - individui di entrambi i gruppi (supporto comune)
#     - solo uomini
#     - solo donne
# - Per le osservazioni che fanno parte del supporto comune si stimano, per
#   ognuno dei due gruppi, le frequenze osservate in ogni strato.
# - Si procede in seguito con la stima dei fattori di riponderazione, che
#   servono a bilanciare la distribuzione congiunta delle caratteristiche di
#   un gruppo a quella dell'altro gruppo. Ci sono due possibilità: bilanciare
#   la distribuzione congiunta delle caratteristiche del gruppo A (uomini) a
#   quella del gruppo B (donne), o viceversa.
# I passaggi appena descritti possono essere eseguiti con la funzione
# reweight_strata_all4 del pacchetto decr:
r01 <- reweight_strata_all4(
  data = salari_inventati,
  treatment = "sesso",
  variables = c("settore_economico", "grado_formazione"),
  y = "salario_mensile",
  weights = "peso"
)

# Prima di effettuare la scomposizione della differenza osservata tra le medie
# salariali di uomini e donne, verifichiamo le medie salariali osservate:
m01 <- margin_mean(r01)
m01
```

```
## # A tibble: 2 x 3
##   sesso   ybar  Nhat
##   <fctr> <dbl> <dbl>
## 1 Uomini  5323 42363
## 2 Donne   3614 27833
```

```
# I salari medi sono nella colonna "ybar", mentre nella colonna "Nhat" c'è una
# stima della numerosità di uomini e donne estratta dai dati.
# Dal campione di dati fittizi, si osserva quindi una differenza tra i salari
# medi di uomini e donne di (in franchi al mese):
margin_difference(m01)
```

```
## [1] 1709.264
```

```
# Adesso possiamo procedere con la scomposizione di questa differenza
# osservata, con i fattori di riponderazione calcolati poco sopra.
# Si tratta della scomposizione in 4 elementi, proposta da Nopo (2008),
# con il metodo che abbiamo proposto nel documento. Prima si suddivide il
# campione nei quattro possibili sottoinsiemi (uomini e donne dentro e fuori
# il supporto comune) e, per ognuno di questi, si stimano alcune statistiche
# che descriviamo sotto:
```

```
s01 <- nopodec_mean(r01)
s01
```

```
## # A tibble: 4 x 6
##   sesso  common_support  ybar ybar_C_A ybar_C_B  Nhat
##   <fctr> <lgl>          <dbl>    <dbl>    <dbl> <dbl>
## 1 Uomini F               6057     6057     6057  4056
## 2 Uomini T               5245     5006     5245 38307
## 3 Donne  F               3369     3369     3369  5496
## 4 Donne  T               3674     3674     4220 22337
```

```
# Per ognuno dei 4 sottoinsiemi sono stati stimati i salari medi ("ybar"),
# i salari medi controfattuali nelle due varianti ("ybar_C_A" e "ybar_C_B") e
# la numerosità ("Nhat").
# A partire da questi elementi, si può procedere alla scomposizione della
# differenza osservata tra le medie salariali di uomini e donne con la funzione
# nopodec del pacchetto decr:
n01_AB <- nopodec(s01, counterfactual = "AB")
unlist(n01_AB)
```

```
##  delta_tot    delta_A    delta_X    delta_S    delta_B
## 1709.26375   77.72983  239.00516 1332.20793   60.32083
```

```
# `delta_tot` è la differenza osservata tra i salari medi degli uomini e delle
# donne (il salario medio osservato degli uomini è superiore a quello delle
# donne). Questa differenza osservata viene suddivisa nelle quattro componenti
# seguenti:
# - `delta_A`: parte della differenza osservata dovuta al fatto che ci sono
#   uomini (gruppo A) che non sono paragonabili, per caratteristiche, a delle
#   donne (gruppo B)
# - `delta_X`: parte spiegata dal fatto che i due gruppi hanno una
#   distribuzione delle caratteristiche diversa
# - `delta_S`: parte non giustificata dalle diverse distribuzioni delle
#   caratteristiche dei due gruppi e potenzialmente dovuta a una differenza
#   nelle strutture remunerative tra i due gruppi
# - `delta_B`: parte della differenza dovuta al fatto che ci sono donne
#   (gruppo B) con delle caratteristiche che nessun uomo (gruppo A) ha.
```

```
# Ora che abbiamo scomposto la differenza tra le medie salariali di uomini e
# donne, possiamo approfondire l'analisi con le scomposizioni delle differenze
# tra i quantili salariali dei due gruppi. Ricordiamo che queste scomposizioni
# sono possibili solo nel supporto comune.
# L'analisi mira in particolare a come si distribuisce la parte non spiegata
# delle differenze salariali, per verificare se è costante oppure se ci sono
# delle differenze tra la parte bassa e quella alta della distribuzione.
# Le funzioni dec_quantile e dec_ del pacchetto decr servono per effettuare le
# scomposizioni delle differenze tra i quantili dei salari.
# Per esempio, la scomposizione della differenza tra il decimo percentile dei
# salari degli uomini e quello delle donne (nel supporto comune) si effettua
# così:
p10 <- dec_quantile(r01, probs = 0.1)
p10_dec <- dec_(p10, counterfactual = "AB")

p10
```

```
## # A tibble: 6 x 7
##   sesso  common_support  yhat yhat_C_A yhat_C_B  Nhat probs
## * <fctr> <lgl>          <dbl>    <dbl>    <dbl> <dbl> <dbl>
## 1 Uomini F               4000     4000     4000  4056 0.100
## 2 Uomini T               3400     3400     3400 38307 0.100
## 3 Donne  F               2300     2300     2300  5496 0.100
## 4 Donne  T               2600     2600     2700 22337 0.100
## 5 Uomini NA              3400       NA       NA 42363 0.100
## 6 Donne  NA              2500       NA       NA 27833 0.100
```

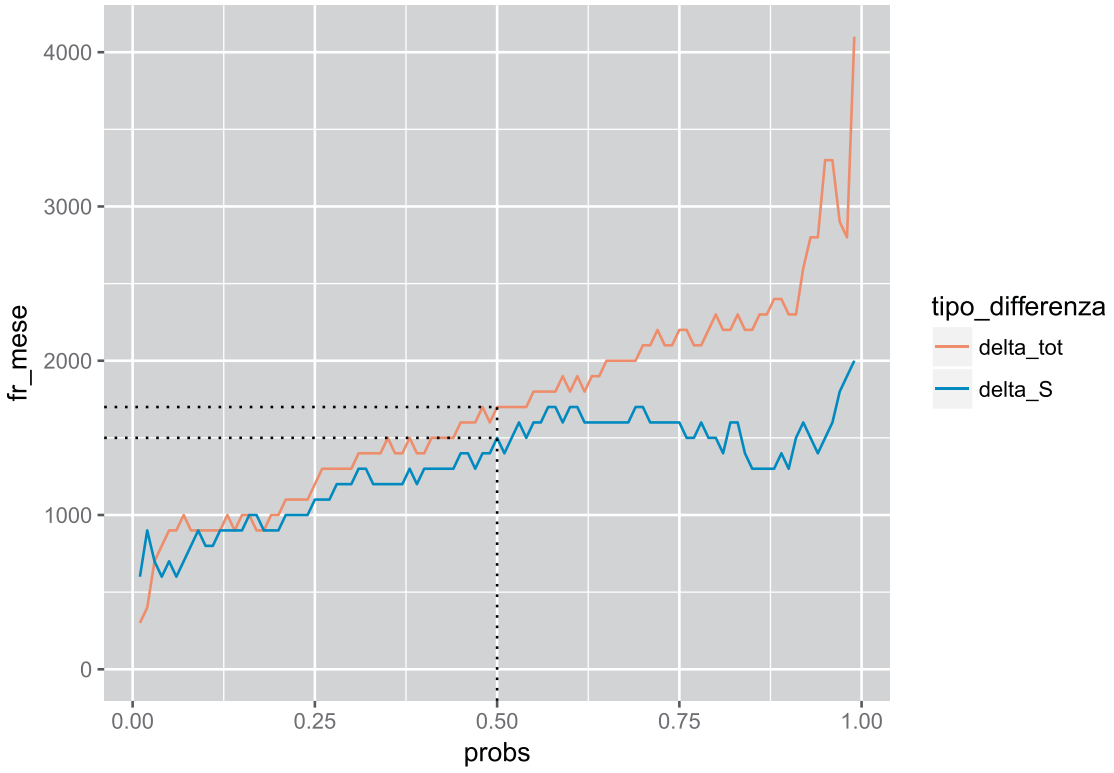
```
unlist(p10_dec)
```

```
##        probs    delta_tot delta_tot_CS     delta_AB      delta_X
##          0.1        900.0        800.0        100.0          0.0
##      delta_S
##        800.0
```

```
# Il decimo percentile dei salari di tutti gli uomini del campione è di 3.400
# franchi al mese, mentre quello delle donne di 2.500 franchi al mese (vedi
# ultime due righe dell'oggetto p10 qui sopra). Ne risulta quindi una
# differenza osservata (complessiva) di 900 franchi mensili (vedi elemento
# delta_tot dell'oggetto p10_dec qui sopra).
# Se si considera solo il supporto comune, il decimo percentile dei salari
# degli uomini è di 3.400 franchi al mese, mentre quello delle donne 2.600
# franchi al mese (vedi seconda e quarta riga dell'oggetto p10 qui sopra).
# La differenza osservata nel supporto comune è quindi di 800 franchi mensili
# (elemento delta_tot_CS dell'oggetto p10_dec qui sopra).
# Questa differenza può essere a sua volta scomposta in una componente spiegata
# dalla diversa distribuzione delle caratteristiche dei due gruppi (il settore
# economico e il grado di formazione, nel nostro esempio) (delta_X dell'oggetto
# p10_dec qui sopra) e in una parte non spiegata (delta_S dell'oggetto p10_dec
# qui sopra). Al livello del decimo percentile dei salari risulta che nel
# supporto comune neanche un franco della differenza osservata può essere
# spiegata dalla diversa distribuzione delle caratteristiche di uomini e donne.
# L'elemento delta_AB dell'oggetto p10_dec è la differenza "residua", imputabile
# al fatto che esistono dei profili di uomini per i quali non esiste nessuna
# donna e profili di donne per i quali non esiste nessun uomo. Si tratta in
# sostanza della somma degli elementi delta_A e delta_B della scomposizione
# della differenza tra le medie salariali presentata prima, che però nel caso
# dei quantili non è scomponibile ulteriormente. Nel nostro esempio, delta_AB
# è pari a 100 franchi al mese.
#
# Con il codice seguente effettuiamo le scomposizioni delle differenze tra i
# percentili dei salari di uomini e donne su vari livelli della distribuzione,
# dal primo al novantanovesimo percentile (con intervalli di un percentile)
```

```
p <- seq(from = 0.01, to = 0.99, by = 0.01)
p01_99 <- lapply(p, function(x) dec_quantile(r01, probs = x))
p01_99_dec <- lapply(p01_99, dec_)
p01_99_dec <- do.call(rbind.data.frame, p01_99_dec)

# Rappresentiamo graficamente la differenza complessiva osservata (delta_tot)
# e la parte non spiegata (delta_S)
library(reshape2) # per la funzione melt
library(ggplot2)  # per le funzioni ggplot, geom_line, annotate e expand_limits
p01_99_dec_melt <- melt(p01_99_dec, id.vars = "probs",
            measure.vars = c("delta_tot", "delta_S"),
            variable.name = "tipo_differenza", value.name = "fr_mese")
g <- ggplot(p01_99_dec_melt,
            aes(x = probs, y = fr_mese, colour = tipo_differenza))
g + geom_line() +
  annotate(geom = "segment", x = -Inf, xend = 0.5, y = 1700, yend = 1700,
           linetype = "dotted") +
  annotate(geom = "segment", x = -Inf, xend = 0.5, y = 1500, yend = 1500,
           linetype = "dotted") +
  annotate(geom = "segment", x = 0.5, xend = 0.5, y = -Inf, yend = 1700,
           linetype = "dotted") +
  expand_limits(y = 0)
```




```
# L'asse delle x del grafico rappresenta i livelli dei quantili (nell'esempio
# sono 99 percentili), mentre sull'asse delle y ci sono le differenze tra i
# percentili dei salari di uomini e donne.
# delta_tot rappresenta, per ogni percentile, la differenza osservata tra i
# salari degli uomini e le donne di tutto il campione. delta_S invece
# rappresenta la differenza tra i percentili dei salari degli uomini e delle
# donne "a parità di caratteristiche", e quindi la parte di differenza non
# spiegata dalla diversa distribuzione delle caratteristiche che i due gruppi
# possono avere.
# Osservando il grafico possiamo fare alcune considerazioni:
# - Le differenze salariali osservate tra uomini e donne variano sui diversi
#   livelli della distribuzione: nella parte bassa le differenze partono da
#   circa 300 franchi mensili (al primo percentile) e arrivano ai 1.700 franchi
#   al livello mediano; nella parte alta della distribuzione i divari tra i
#   percentili salariali degli uomini e delle donne raggiungono 4.100 franchi
#   (all'estremo del novantanovesimo percentile).
# - Nella parte bassa della distribuzione dei salari, le differenze osservate
#   e le differenze "a parità di caratteristiche" (non spiegate) non mostrano
#   delle grandi differenze (da -100 a 300 franchi).
# - A partire circa dal settantesimo percentile, invece, la parte non spiegata
#   si distanzia sempre di più dalla differenza osservata totale, pur
#   mantenendo un andamento piuttosto crescente.
# - Riassumendo, anche la parte non spiegata delle differenze salariali mostra
#   un andamento tendenzialmente crescente con il crescere del livello dei
#   percentili, ma in modo molto meno marcato rispetto alle differenze
#   osservate senza tenere conto delle diversità in termini di caratteristiche
#   di uomini e donne. La parte non spiegata parte da circa 600 franchi al
#   livello più basso considerato nell'analisi (il primo percentile) fino a
#   raggiungere i 2.000 franchi al livello più alto considerato (il
#   novantanovesimo percentile).
```

APPENDICE B

Formule degli stimatori utilizzati

B.1 Stimatore degli effettivi (numero di salariati o di posti di lavoro)

La stima degli effettivi viene fatta semplicemente sommando i pesi di campionamento:

$$\widehat{N} = \sum_{i=1}^{n} W_i.$$

Nel caso di stime di effettivi di sottoinsiemi della popolazione di riferimento, si utilizza la stessa formula, utilizzando l'indicazione di appartenenza al sottoinsieme di interesse come indice della sommatoria, oppure all'interno di una funzione indicatrice. Per esempio, lo stimatore del numero di individui del gruppo A risulta:

$$\widehat{N}_A = \sum_{i \in A} W_i = \sum_{i=1}^{n} W_i \cdot \mathbf{1}(i \in A).$$

B.2 Funzione di ripartizione empirica dei salari

A partire da un certo livello di salari, y, stima la proporzione di salari nel campione inferiore o uguale a questo livello:

$$\widehat{F}_Y(y) = \frac{\sum_{i=1}^{n} W_i \cdot \mathbf{1}(Y_i \leq y)}{\sum_{i=1}^{n} W_i}.$$

B.3 Quantili dei salari (mediana, percentili, ...)

I quantili empirici non sono unici. Per questo motivo, ci sono vari metodi per stimarli. Qui utilizziamo quello utilizzato dall'Ufficio federale di statistica nell'ambito della RSS (vedi Graf (2002), pag. 30).

Siano $Y_{[1]} \leq Y_{[2]} \leq \ldots \leq Y_{[n]}$ i salari del campione ordinati dal più piccolo al più grande. Il quantile empirico di ordine p, con p compreso tra 0 e 1, è definito tramite l'inversione della funzione di ripartizione empirica dei salari:

$$\widehat{F}^{-1}(p) = \begin{cases} \frac{1}{2}(Y_{[i]} + Y_{[i+1]}) & \text{se } \widehat{F}_Y(Y_{[i]}) = p \\ Y_{[i+1]} & \text{se } \widehat{F}_Y(Y_{[i]}) < p < \widehat{F}_Y(Y_{[i+1]}). \end{cases}$$

B.4 Salario medio

- Salario medio osservato del gruppo A:

$$\widehat{\overline{Y}}_A = \frac{\sum_{i \in A} W_i \cdot Y_i}{\sum_{i \in A} W_i}.$$

- Salario medio osservato del gruppo B:

$$\widehat{\overline{Y}}_B = \frac{\sum_{i \in B} W_i \cdot Y_i}{\sum_{i \in B} W_i}.$$

- Differenza osservata tra le medie salariali dei gruppi A e B:

$$\widehat{\Delta}^{\mu}_{O} = \widehat{\overline{Y}}_A - \widehat{\overline{Y}}_B.$$

- Differenza tra le medie salariali nel **supporto comune**:

$$\widehat{\Delta}^{\mu}_{O|\mathcal{S}} = \widehat{\overline{Y}}_{A|\mathcal{S}} - \widehat{\overline{Y}}_{B|\mathcal{S}}.$$

- Il supporto comune, $\mathcal{S}$, è l'insieme delle combinazioni di caratteristiche in cui si osservano individui di entrambi i gruppi. Il suo complemento, $\mathcal{S}^C$, è l'insieme delle combinazioni di caratteristiche in cui si osservano solo individui di un gruppo e nessuno dell'altro.

- Salario medio del gruppo A osservato nel **supporto comune**:

$$\widehat{\overline{Y}}_{A|\mathcal{S}} = \frac{\sum\limits_{\substack{i \in A \\ X_i \in \mathcal{S}}} W_i \cdot Y_i}{\sum\limits_{\substack{i \in A \\ X_i \in \mathcal{S}}} W_i}.$$

In pratica si tratta della media dei salari degli individui del gruppo A ($i \in A$) che hanno caratteristiche che appartengono al supporto comune ($X_i \in \mathcal{S}$).

APPENDICE C

R session info

```
sessionInfo()
```

```
## R version 3.4.3 (2017-11-30)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Manjaro Linux
##
## Matrix products: default
## BLAS: /usr/lib/libblas.so.3.8.0
## LAPACK: /usr/lib/liblapack.so.3.8.0
##
## locale:
##  [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
##  [3] LC_TIME=en_US.UTF-8        LC_COLLATE=en_US.UTF-8
##  [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
##  [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
##  [9] LC_ADDRESS=C               LC_TELEPHONE=C
## [11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
## [1] grid      stats     graphics  grDevices utils     datasets  methods
## [8] base
##
## other attached packages:
## [1] ggplot2_2.2.1   reshape2_1.4.3  decr_0.0.7.9031
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_0.12.14     highr_0.6        pillar_1.0.1     compiler_3.4.3
##  [5] plyr_1.8.4       bindr_0.1        tools_3.4.3      boot_1.3-20
##  [9] digest_0.6.13    evaluate_0.10.1  tibble_1.4.1     gtable_0.2.0
## [13] pkgconfig_2.0.1  rlang_0.1.6      cli_1.0.0        yaml_2.1.16
## [17] bindrcpp_0.2     xml2_1.1.1       httr_1.3.1       stringr_1.2.0
## [21] dplyr_0.7.4      knitr_1.18       rprojroot_1.3-2  tidyselect_0.2.3
## [25] glue_1.2.0       R6_2.2.2         tidyr_0.7.2      purrr_0.2.4
## [29] magrittr_1.5     backports_1.1.2  scales_0.5.0     assertthat_0.2.0
## [33] mime_0.5         colorspace_1.3-2 labeling_0.3     utf8_1.1.3
## [37] stringi_1.1.6    lazyeval_0.2.1   munsell_0.4.3    crayon_1.3.4
```

BIBLIOGRAFIA

Blinder, Alan S. 1973. "Wage Discrimination: Reduced Form and Structural Estimates." *The Journal of Human Resources* VIII (4):436–55.

DiNardo, John, Nicole M. Fortin e Thomas Lemieux. 1996. "Labor Market Institutions and the Distribution of Wages, 1973-1992: A Semiparametric Approach." *Econometrica* 64 (5):1001–44.

Fortin, Nicole M., Thomas Lemieux e Sergio Firpo. 2011. "Decomposition Methods in Economics." In, edited by O. Ashenfelter and D. Card, 4A:1–102. Handbook of Labor Economics. Elsevier.

Graf, Monique. 2002. "Enquête Suisse Sur La Structure Des Salaires 2000. Plan d'échantillonage, Pondération et Méthode d'estimation Pour Le Secteur Privé." Rapport de méthode 338-0010. Office fédéral de la statistique. *https://www.bfs.admin.ch/bfs/fr/home/services/recherche/rapports-methodologiques.assetdetail.344403.html*.

Iacus, Stefano M., Gary King e Giuseppe Porro. 2011. "Causal Inference Without Balance Checking: Coarsened Exact Matching." *Political Analysis* 20:1–24.

Nopo, Hugo. 2008. "Matching as a Tool to Decompose Wage Gaps." *Review of Economics and Statistics* 90 (2):290–99.

Oaxaca, Ronald. 1973. "Male-Female Wage Differentials in Urban Labor Markets." 14 (3):693–709.

Impaginazione:
Sharon Fogliani

Ufficio di statistica

Repubblica e Cantone Ticino
Dipartimento delle finanze e dell'economia
Divisione delle risorse

Febbraio 2018

Ufficio di statistica
Via Bellinzona 31
6512 Giubiasco

+41 (0)91 814 50 11
dfe-ustat@ti.ch
www.ti.ch/ustat