RICCHEZZA E ISTRUZIONE (NEL MONDO)

Descrizione della base dati

Il dataste considerato è tratto dal report annuale dell'UNESCO, relativo a 102 Stati, sui quali sono rilevate un certo numero di caratteristiche, alcune delle quali di natura socio-economica (ad esempio, prodotto nazionale lordo, speranze di vita, tassi di natalità e mortalità, mortalità infantile, tassi di urbanizzazione), altre relative al sistema istruzione del Paese (tra esse, la spesa pubblica per l'istruzione, il numero di individui che accedono all'istruzione superiore, la percentuale di donne nell'ambito del personale docente).

Le variabili rilevate sono le seguenti:

I GRUPPO (indicatori socio-economici e demografici)

MACROAREA (gruppo o area socio-economica di appartenenza)
POPOLAZIONE (espressa in migliaia di abitanti)
DENSITA' DI POPOLAZIONE (espressa in abitanti per chilometro quadrato)
TASSO DI CRESCITA (espresso in aumento annuo percentuale della popolazione)
TASSO DI NATALITA' (espresso in nascite per 1000 abitanti)
TASSO DI MORTALITA' (espresso in morti per 1000 abitanti)
TASSO DI FECONDITA' (espresso in numero medio di figli per donna)
TASSO DI URBANIZZAZIONE (espresso in percentuale degli abitanti che vivono in città)
SPERANZA DI VITA MASCHILE
SPERANZA DI VITA FEMMINILE
MORTALITA' INFANTILE (espresso in numero di bambini morti per 1000 nati vivi)
PRODOTTO INTERNO LORDO PRO CAPITE (in dollari)

II GRUPPO (indicatori istruzione ed università)

TASSO DI ALFABETIZZAZIONE (espresso in percentuale di abitanti che sanno leggere e scrivere)
TASSO DI ALFABETIZZAZIONE MASCHILE
TASSO DI ALFABETIZZAZIONE FEMMINILE
TASSO DI ISTRUZIONE SUPERIORE (espresso in iscritti a corsi universitari per 1000 abitanti)
SPESE PUBBLICHE PER ISTRUZIONE (percentuale del PIL destinato a spese per istruzione)
RAPPORTO STUDENTI-DOCENTI (numero di studenti per docente)
PRESENZA FEMMINILE NELL'INSEGNAMENTO (percentuale di femmine tra i docenti)
PATRIMONIO LIBRARIO UNIVERSITARIO (numero testi posseduti dalle biblioteche universitarie)
BIBLIOTECHE UNIVERSITARIE (numero delle biblioteche universitarie)

Intenti dell'analisi

Ci si propone di analizzare questi dati, con l'obiettivo di investigare diversi aspetti, con particolare riguardo alle ***interazioni tra parametri socio-economici e culturali***.

L'obiettivo è duplice.

Da una parte ci si propone di comprendere come le condizioni socio-economiche e demografiche di un Paese possano variare in contesti differenti e essere tra loro legate. Ad esempio, con riferimento al primo gruppo di caratteristiche, ci si chiede quanto la ricchezza sia distribuita equamente tra gli Stati in esame e all'interno di macroaree definite, e quanto il PIL, indicatore globale di ricchezza, sia legato ad alcune caratteristiche demografiche. Inoltre, è interessante comprendere quali differenze sussistano relativamente alle condizioni di vita di maschi e femmine, e come queste differenze varino al variare del gruppo di appartenenza dello Stato. Oppure, come variabili che indicano il livello di salute di uno Stato (quali la speranza di vita e la mortalità infantile) sono legate ad indicatori globali di ricchezza.

Per ciò che riguarda il secondo gruppo di variabili, si è interessati a valutare se e come l'accesso e la qualità dell'istruzione siano legati ad indicatori di ricchezza o ad altre variabili di tipo socio-economico o demografico. In particolare, il numero di studenti per docente può essere interpretato come indicatore di efficienza del sistema istruzione e di buone condizioni sociali? La presenza femminile nell'insegnamento è legata ad altri indicatori della condizione femminile nel Paese?

In particolare proponiamo una serie di domande, alcune delle quali saranno analizzate nel seguito. L'obiettivo è quello di rispondere ad esse traducendo la questione in termini statistici, usando gli strumenti più idonei, rilevando problemi o difficoltà di interpretazione nei risultati.

1. ***Com'è distribuita la ricchezza tra i vari paesi? Come varia la distribuzione della ricchezza tra le diverse macroaree?***

Per rispondere a questa domanda, si possono in primo luogo determinare rappresentazioni grafiche opportune per il PIL, quindi calcolare indicatori di variabilità e concentrazione per lo stesso carattere. Inoltre, un'analisi stratificata, per macroarea, si può effettuare utilizzando le distribuzioni subordinate e le corrispondenti misure di sintesi.

Concetti richiamati e strumenti statistici utilizzati: Misure di sintesi (posizione e dispersione). Rappresentazioni grafiche: box-plot; istogramma (scelta delle classi); curva di concentrazione, indice di concentrazione di Gini. Analisi stratificata per macroarea: box-plot affiancati, misure di sintesi subordinate. I valori estremi

2. ***Come varia tra i Paesi la speranza di vita alla nascita? Ci sono caratteristiche diverse tra le distribuzioni delle speranze di vita dei maschi e delle femmine?***

Anche in questo caso, è opportuno determinare e confrontare indicatori di sintesi per le due variabili in esame, anche nell'ambito di ciascuna macroarea.

Concetti richiamati e strumenti statistici utilizzati: Misure di sintesi (posizione e dispersione). Analisi delle relazioni tra due caratteri: diagramma di dispersione; coefficiente di correlazione (non-robustezza). Analisi stratificata per macroarea: medie condizionate; coefficiente di variazione condizionati

3. ***I tassi di natalità e mortalità sono tra loro associati? In che modo? A quali altre variabili sono significativamente legati?***

Per valutare l'associazione tra le due variabili, si possono determinare opportuni indici (coefficiente di correlazione lineare) e riportare il grafico di dispersione. Inoltre, si può discutere la scelta della curva che meglio descrive la dipendenza tra le due variabili.

Concetti richiamati e strumenti statistici utilizzati: Analisi delle relazioni tra due caratteri: diagramma di dispersione; coefficiente di correlazione (non-robustezza); misure di concordanza – tau di Kendall e indice di Spearman. Interpolazione di una nuovola di punti (previsione/spiegazione di un carattere in funzione di un altro): retta di regressione, indice di determinazione, funzioni polinomiali

4. ***Come cresce (o decresce) la popolazione? Come variano gli indicatori d'incremento demografico nelle diverse macroaree?***

Oltre ad analizzare singolarmente, mediante la determinazione di misure di sintesi e grafici, le variabili "Tasso d'incremento demografico" e "Tasso di fecondità", si possono effettuare analisi di associazione tra esse e discutere la regressione di queste variabili sul PIL.

Concetti richiamati e strumenti statistici utilizzati: Analisi stratificata per macroarea: medie e mediane condizionate. Interpolazione di una nuovola di punti (previsione/spiegazione di un carattere in funzione di un altro): retta di regressione, indice di determinazione, funzioni polinomiali; valutazione dell'impatto della variabile esplicativa.

5. ***L'accesso all'istruzione è simile nelle varie macroaree? A quali altre variabili socio-economiche sono legati indicatori quali tassi di alfabetizzazione e accesso all'istruzione superiore? Sono tra loro associati questi due indicatori? In che modo? Vi sono differenze rilevanti tra maschi e femmine?***

Concetti richiamati e strumenti statistici utilizzati: Misure di sintesi (posizione e dispersione). Valori mancanti. Analisi stratificata per macroarea: medie condizionate. Analisi delle relazioni tra due caratteri: diagramma di dispersione; coefficiente di correlazione (non-robustezza); misure di concordanza – tau di Kendall e indice di Spearman. Interpolazione di una nuovola di punti (previsione/spiegazione di un carattere in funzione di un altro): retta di regressione, indice di determinazione, funzioni esponenziali. Analisi stratificata per macroarea dei coefficienti di correlazione

6. ***Cosa si può dire, in generale, sulla differenza tra condizione maschile e condizione femminile rispetto ai diversi aspetti considerati nell'analisi?***

Le differenze tra tassi di alfabetizzazione e tra attese di vita forniscono indicazioni sulla condizione femminile. La variazione di queste differenze al variare del PIL ed al variare della macroarea informano su quanto la condizione femminile sia legata al livello di benessere complessivo del paese.

Concetti richiamati e strumenti statistici utilizzati: Misure di sintesi (posizione e dispersione). Valori mancanti. Rappresentazioni grafiche: box-plot. Analisi delle relazioni tra due caratteri: diagramma di dispersione; coefficiente di correlazione (non-robustezza); misure di concordanza – tau di Kendall e indice di Spearman. Interpolazione di una nuovola di punti (previsione/spiegazione di un carattere in funzione di un altro): retta di regressione, indice di determinazione, split del modello

7. ***Le spese che lo Stato destina all'istruzione (spesa in percentuale sul PIL) sono positivamente associate alla ricchezza del Paese (descritta dallo stesso PIL)? Il tipo e grado di associazione è diverso nelle varie macroaree?***

L'analisi, anche stratificata, della distribuzione della spesa destinata all'istruzione è il primo passo. Anche in questo caso, la regressione di questa variabile su altre variabili relative alla situazione complessiva di benessere del paese, come il PIL, forniscono informazioni sull'interdipendenza di istruzione e benessere.

Concetti richiamati e strumenti statistici utilizzati: Misure di sintesi (posizione e dispersione). Valori mancanti. Rappresentazioni grafiche: box-plot, istogramma. Analisi stratificata per macroarea: box plot affiancati. Interpolazione di una nuovola di punti (previsione/spiegazione di un carattere in funzione di un altro): retta di regressione, indice di determinazione

Altre domande di interesse:

Il tasso di urbanizzazione varia significativamente nelle diverse macroaree? E' direttamente legato ad altre caratteristiche socio-economiche? Come varia al variare della popolazione?

La percentuale di docenti di sesso femminile è legata ad altri indicatori della condizione femminile, già descritti in precedenza? Come ed in che misura?

Un indicatore di efficienza del sistema istruzione potrebbe essere fornito dal legame tra spese pubbliche per istruzione e numero di studenti che usufruiscono dell'istruzione terziaria. E' effettivamente così? Cosa si può dedurre da questa analisi?

Il numero di studenti per docente è, a sua volta, un indicatore di efficienza e bontà del sistema istruzione? Perché ed in che misura?

Indicazioni tecniche

Alcune indicazioni per la lettura di questo documento:

- breve spiegazione se una parola è evidenziata in verde, questo simbolo alla fine o all'inizio della riga in cui si trova la parola contiene una brevissima spiegazione del concetto statistico. Per leggere il contenuto del commento basta toccare il simbolo con il mouse.

- [*Commento o descrizione più dettagliata di una tecnica*] una frase evidenziata in giallo rimanda ad un documento che contiene alcune considerazioni o metodologiche o sulle modalità di utilizzo e di interpretazione degli strumenti statistici cui si fa riferimento. Per aprire il documento, selezionate la nella barra del menu di Adobe:

Cliccando con il mouse sulla frase, si aprirà il documento.

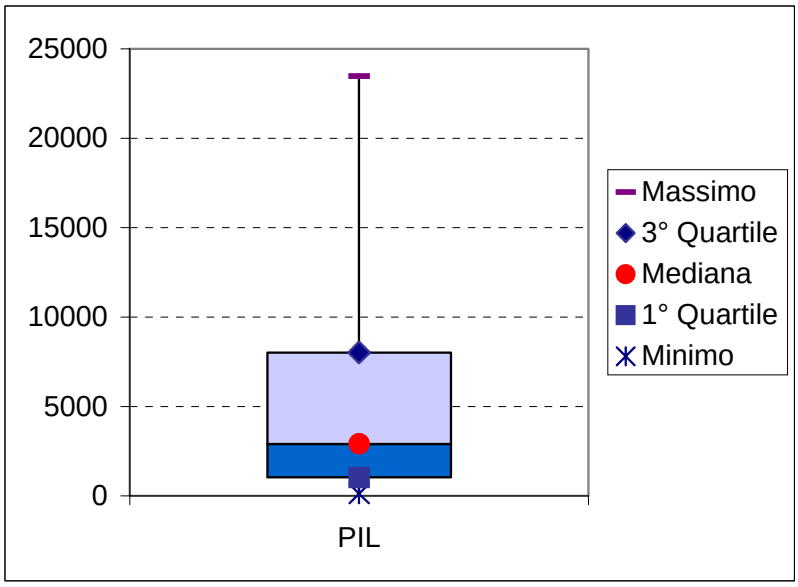
NB: i **richiami metodologici** cui si rimanda nel testo **(ordinati per argomento)** sono tutti contenuti in questo documenti [*Richiami metodologici*]

1. *Come è distribuita la ricchezza tra i vari paesi? Come varia la distribuzione della ricchezza tra le diverse macroaree?*

La variabile più idonea a descrivere la ricchezza di un Paese è ovviamente il Prodotto Interno Lordo (pro capite). Una prima analisi elementare[1] per comprendere come la ricchezza sia distribuita nel mondo si può effettuare rilevando alcuni indicatori sintetici (di posizione e variabilità) di questa variabile, quali media aritmetica, mediana e quartili, massimo e minimo, scarto quadratico medio (o deviazione standard).

Media	6006,108
Mediana	2912
Moda	1500
Varianza	43852813
Deviazione standard	6622,146
Minimo	122
Massimo	23474
Somma	612623
Conteggio	102

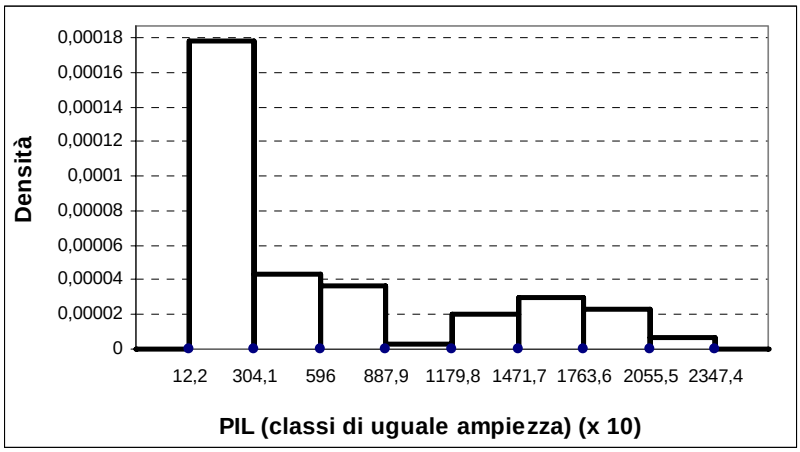
Dalla tabella (ottenuta con EXCEL) si evince che il prodotto interno lordo pro capite è, in media, 6006.11 dollari, mentre la sua mediana è pari a 2912 dollari; osservando anche i valori dei quartili e del minimo e massimo, si può concludere che la distribuzione del prodotto interno lordo è asimmetrica, più precisamente obliqua destra. Questi valori sono riassunti in un grafico, il boxplot, che, fornendo una rappresentazione sintetica della forma della distribuzione [*Box plot: una rappresentazione sintetica della distribuzione*] conferma le considerazioni che su essi si basano. Il boxplot[2] ha l'aspetto seguente:



Se il box-plot riesce a dare una rappresentazione sintetica della distribuzione, l'istogramma permette invece un maggior dettaglio. Uno dei problemi principali nel costruire questo tipo di grafico è però la scelta delle classi da utilizzare nella rappresentazione, sia per quanto riguarda il numero di classi sia per quanto riguarda gli estremi delle classi stesse. Un semplice criterio consiste nell'utilizzare classi di uguale ampiezza. In questo caso però, il box plot evidenzia una distribuzione fortemente asimmetrica, e la scelta di questo tipo di classi non è indicata.

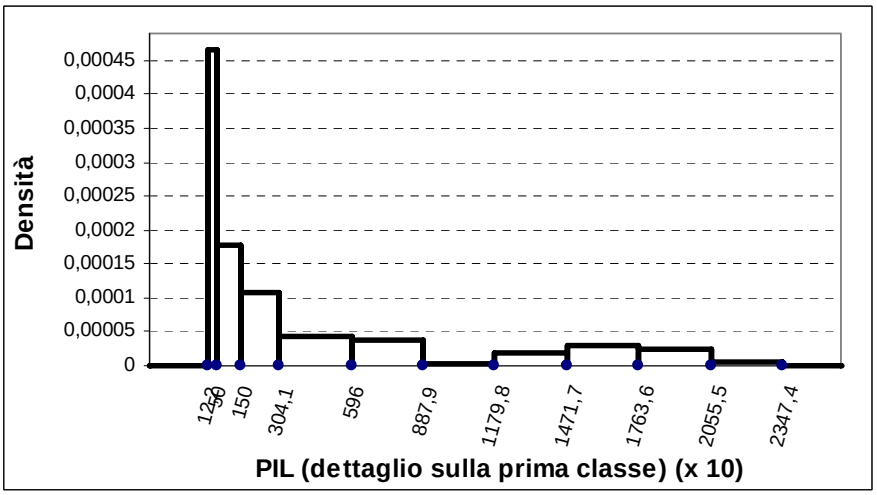
[1] Risultato ottenuto utilizzando Excel (Stumenti, Analisi dei dati, Statistica descrittiva).
[2] Risultato ottenuto utilizzando EXCEL (macro Stat4038).

Riportiamo infatti di seguito l'istogramma relativo alla distribuzione del prodotto interno lordo scegliendo 8 classi di uguale ampiezza.



Notiamo che questo istogramma[3] "descrive" in modo molto dettagliato la lunga coda destra della distribuzione ma non fornisce una descrizione adeguata della classe più rilevante (il 50% degli stati ha un PIL pro capite compreso tra 122 e 304 dollari). Decidiamo quindi di suddividere la prima classe in sottoclassi.

Dopo diversi tentativi, otteniamo che la rappresentazione che meglio rappresenta il carattere è quella in cui la prima classe è suddivisa in 3 sottoclassi cui compete almeno approssimativamente la stessa frequenza (circa 18 stati). [*Istogramma: cautele nella scelta delle classi*]

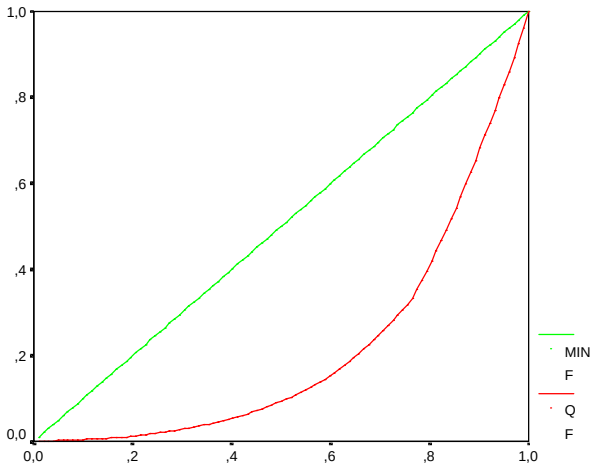


Box plot e istogramma evidenziano entrambi la presenza di un numero rilevante di Stati poveri (con basso PIL pro capite) e frequenze via via più ridotte di stati più ricchi. Anche gli indicatori di variabilità, con il coefficiente di variazione superiore a 1 (in base ai dati della tabella, il coefficiente di variazione è pari a 6622.146/6006.108=1.103), indicano un notevole disuguaglianza nella distribuzione della ricchezza.

Inoltre, le considerazioni effettuate riguardo l'asimmetria della distribuzione del PIL pro-capite indicano come, quale misura di posizione centrale della distribuzione, sia più opportuno utilizzare la mediana rispetto alla media aritmetica (e, quindi, sintetizzare la variabile in esame col valore 2912). [*Media, mediana, media troncata per distribuzioni asimmetriche*]

La curva di concentrazione riportata sotto mostra quale sia il livello di disuguaglianza nella distribuzione del prodotto interno lordo; essendo la curva in posizione all'incirca intermedia tra asse delle ascisse (caso della concentrazione massima) e bisettrice del I quadrante (situazione di concentrazione minima), si può ritenere di media intensità il livello di concentrazione. Il calcolo di un indicatore specifico quale il rapporto di concentrazione di Gini (pari a 0.576) conferma e precisa questa indicazione.

[3] Risultato ottenuto utilizzando EXCEL (macro Stat4038).

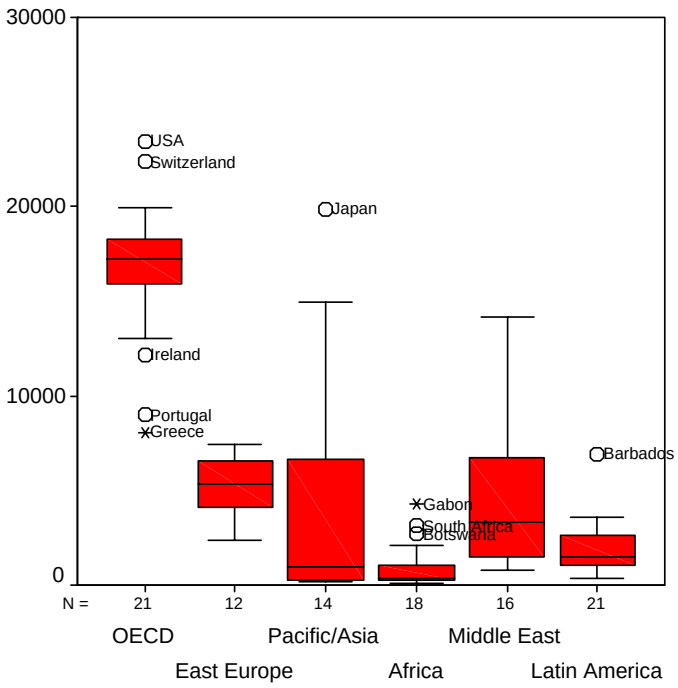


Volendo entrare più nel dettaglio, ci chiediamo se la distribuzione della ricchezza dipende dalla macroarea. Un modo per farlo è quello di procedere ad un'analisi stratificata. [*Analisi stratificata*] Riportiamo quindi di seguito i box-plot affiancati [*Analisi stratificata: box plot affiancati*].
Per un più agevole confronto tra le distribuzioni stratificate determiniamo anche indicatori di posizione e variabilità del PIL area per area e confrontiamo questi valori.
[*Analisi stratificata: le misure di sintesi*]

Case Summaries

Prodotto interno lordo pro capite

Macroarea (regione o	N	Mean	Median	Minimum	Maximum	Std. Deviation	Variance
OECD	21	16610,86	17245,00	8060	23474	3725,97	1,4E+07
East Europe	12	5152,33	5368,00	2340	7400	1625,82	2643292
Pacific/Asia	14	4628,93	933,50	202	19860	6749,49	4,6E+07
Africa	18	1022,33	383,00	122	4283	1207,77	1458704
Middle East	16	4800,56	3360,50	748	14193	4136,94	1,7E+07
Latn America	21	1997,67	1500,00	383	6950	1482,12	2196689
Total	102	6006,11	2912,00	122	23474	6622,15	4,4E+07



La tabella ed il grafico mettono in luce alcune caratteristiche interessanti; in primo luogo, ovviamente, si osserva come i Paesi dell'OECD (ovvero, i paesi dell'Europa occidentale, del nord America, l'Australia e la Nuova Zelanda) abbiano livelli di ricchezza nettamente superiori a quelli degli altri 5 gruppi (le medie e mediane lo indicano chiaramente). In particolare, i valori più bassi del PIL pro capite per i paesi di questa area (osservati per Grecia, Portogallo e Irlanda) risultano comunque superiori ai valori massimi osservati per altre macroaree (Europa dell'est, Africa e America Latina).
Inoltre, la distribuzione del PIL è omogenea (scarsamente variabile) e sostanzialmente simmetrica in alcuni gruppi, mentre in altri no. Ad esempio, nel primo gruppo (paesi OECD) c'è una sostanziale simmetria (lo si evince dai valori di media, mediana, quartili e dal boxplot) ed una non elevata variabilità (il coefficiente di variazione, pari a 3725.9/16610.86=0.224, è di molto inferiore a quello generale che, come già osservato, vale 1.103), il che indica buona omogeneità, rispetto alla ricchezza, dei paesi di questo gruppo. [*Coefficiente di variazione e scarto quadratico medio*]
Al contrario, il terzo gruppo (Asia-Pacifico) presenta forte variabilità e forte asimmetria (verso destra): in un gruppo con la gran parte dei paesi sostanzialmente poveri, vi sono pochi stati con PIL elevato (in particolare, il Giappone viene identificato come outlier nel grafico).
[*Valori estremi e outliers*]

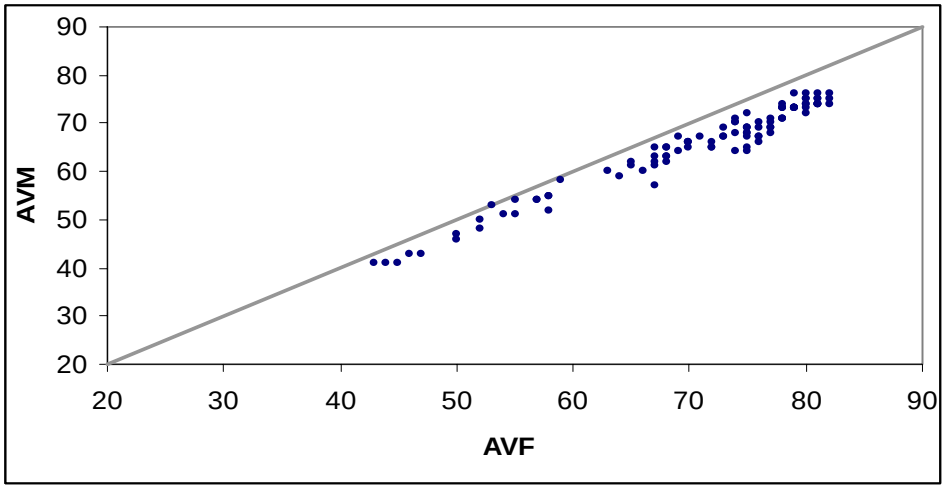
2. *Come varia tra i Paesi la speranza di vita alla nascita? Ci sono caratteristiche diverse tra le distribuzioni delle speranze di vita dei maschi e delle femmine?*

La speranza di vita alla nascita indica sostanzialmente quanti anni un nato può attendersi di vivere, se le condizioni socio-economiche e di benessere sanitario rimanessero invariate.

Iniziamo col fornire indicatori di sintesi per le variabili d'interesse (in questo caso le speranze di vita dei maschi e delle femmine):

	SPERANZA FEMMINE	SPERANZA MASCHI
Media	70,55882	65,20588
Mediana	74	67
Moda	75	73
Varianza	103,1599	81,80868
Deviazione standard	10,15676	9,044815
Minimo	43	41
Massimo	82	76
Somma	7197	6651

Le tabelle evidenziano come, mediamente, le speranze di vita siano superiori per le femmine che per i maschi, e che, allo stesso tempo, ci sia una certa variabilità tra Paesi.
Se a questo punto rappresentiamo in un grafico di dispersione i valori di speranza di vita per i maschi e femmine per i Paesi osservati, ecco che altri elementi vengono in luce.



Innanzi tutto, vediamo che le coppie di valori osservati giacciono tutti sotto la bisettrice del primo quadrante: il fatto che la speranza di vita delle femmine sia più alta della speranza di vita dei maschi non è vero solo "in media", ma vale per tutti i paesi osservati. i due caratteri sono associati in modo diretto (concordante) e quasi perfettamente lineare: la cosa è confermata dal calcolo del coefficiente di correlazione lineare, il cui valore, come si vede dalla tabella che segue, è 0.981.

	AVM	AVF
AVM	1	
AVF	**0,98144**	1

Si osservi che l'assenza di outliers nella distribuzione congiunta delle variabili AVM e AVF e le caratteristiche del grafico (che mostra, come già detto, una quasi perfetta linearità) indicano quanto il valore del coefficiente di correlazione lineare sia affidabile.
[*Cautele nella valutazione del coefficiente di correlazione lineare e di determinazione*]
Questa sostanziale concordanza tra le speranze di vita maschili e femminili è assolutamente comprensibile (e ce l'aspettavamo!): la speranza di vita è in un certo senso un indicatore delle

condizioni di vita del Paese. Quindi, ci possiamo attendere che dove le condizioni di vita peggiorano per i maschi, questo avvenga anche per le femmine e viceversa.

Per comprendere se parte di questa variabilità è imputabile all'appartenenza dei vari paesi a macroaree diverse (che possiamo pensare più omogenee per condizioni di vita) può essere utile procedere con un'analisi stratificata. [*Analisi stratificata: le misure di sintesi*]

Case Summaries

Macroarea (regione o gruppo socio-economico)		Attesa di vita media per femmine	Attesa di vita media per maschi
OECD	N	21	21
	Mean	80,10	73,71
	Median	80,00	74,00
	Std. Deviation	1,18	1,15
East Europe	N	12	12
	Mean	75,75	67,25
	Median	76,00	67,50
	Std. Deviation	,97	2,09
Pacific/Asia	N	14	14
	Mean	69,00	64,64
	Median	70,50	65,50
	Std. Deviation	9,25	7,56
Africa	N	18	18
	Mean	54,78	51,17
	Median	55,00	51,50
	Std. Deviation	7,88	7,28
Middle East	N	16	16
	Mean	71,69	67,44
	Median	72,50	67,00
	Std. Deviation	4,63	4,03
Latn America	N	21	21
	Mean	71,76	66,24
	Median	75,00	68,00
	Std. Deviation	7,39	7,33
Total	N	102	102
	Mean	70,56	65,21
	Median	74,00	67,00
	Std. Deviation	10,16	9,04

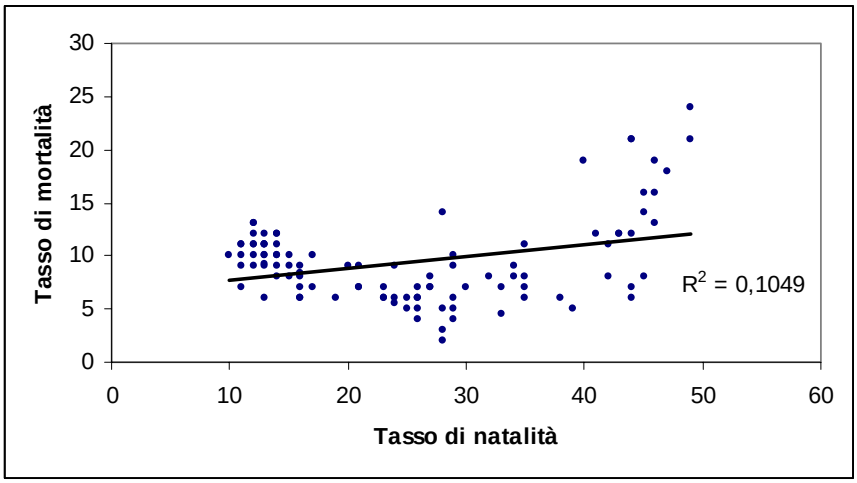
Utilizzando le medie e gli scarti quadratici medi riportati nella tabella, si possono naturalmente calcolare anche i coefficienti di variazione [*Coefficiente di variazione e scarto quadratico medio*] delle attese di vita (maschile e femminile) nelle diverse macroaree; li riportiamo nella tabella che segue.

Macroarea	OECD	Est Europa	Asia Pacifico	Africa	Medio oriente	America Latina
Coeff. Variazione AVF	0.015	0.013	0.134	0.144	0.064	0.103
Coeff. Variazione AVM	0.016	0.031	0.117	0.142	0.060	0.111

Effettivamente la diversità tra le attese di vita medie nelle differenti macroaree è notevole; in particolare, risulta evidente la differenza tra paesi OECD e Africa. Ancora, i coefficienti di variazione mettono in luce variabilità molto superiore, rispetto alle caratteristiche in esame, nelle aree Asia-Pacifico, Africa e America Latina rispetto alle aree OECD e Europa orientale. Un'ultima osservazione: nell'Europa dell'est la variabilità è notevolmente superiore per l'attesa di vita dei maschi rispetto a quella delle femmine, il che indica maggiore omogeneità di questi paesi rispetto alla seconda variabile.

3. *I tassi di natalità e mortalità sono tra loro associati? In che modo? A quali altre variabili sono significativamente legati?*

Il grafico di dispersione relativo a "Tasso di natalità" e a "Tasso di mortalità", ed il corrispondente coefficiente di correlazione lineare, pari a 0.324, forniscono una prima valutazione sul tipo e livello di associazione tra le due variabili:



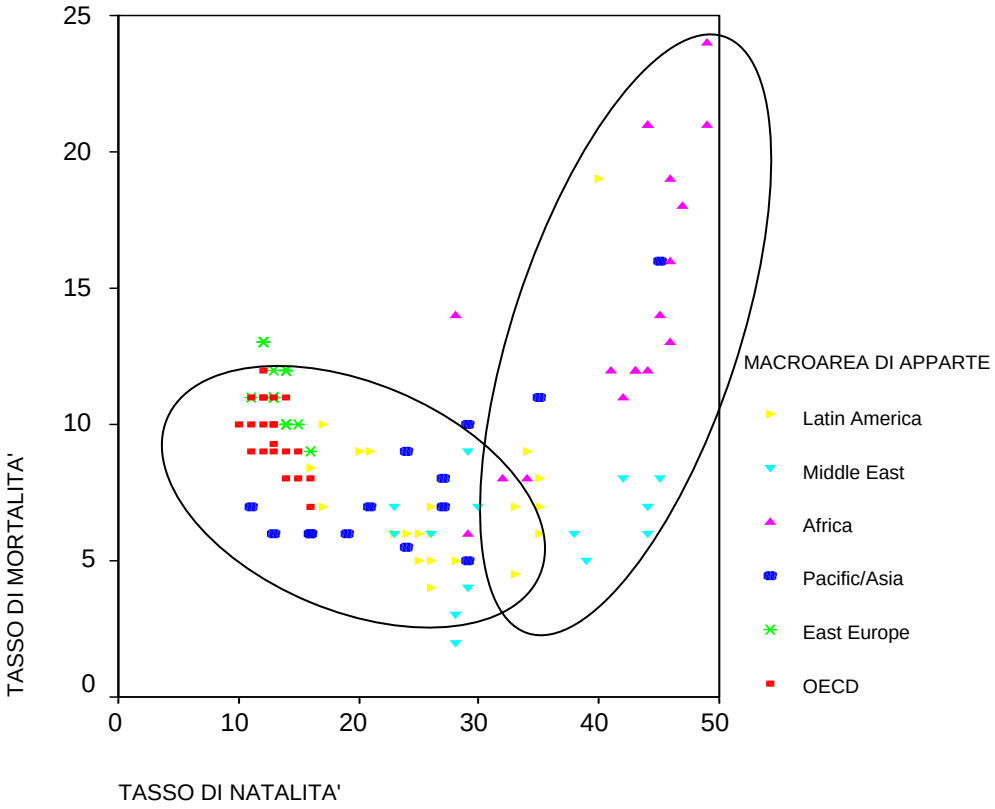
Vi è una (debole) associazione lineare diretta tra i due caratteri; quindi, la retta dei minimi quadrati non spiega adeguatamente la dipendenza del tasso di mortalità dal tasso di natalità. In particolare, il coefficiente di determinazione, riportato nel grafico e pari a 0.1049 evidenzia che la retta non è in grado di spiegare una percentuale soddisfacente della variabilità della variabile dipendente.
Naturalmente, il fatto che l'associazione lineare tra i due caratteri sia debole non implica che i due caratteri non siano legati da un altro tipo di relazione.
Cerchiamo in primis misure di sintesi più adeguate del coefficiente di correlazione lineare per descrivere l'associazione tra i due caratteri. Piuttosto che indagare sulla forza del legame lineare, cerchiamo di valutare l'esistenza di una generica relazione di *concordanza*: due caratteri si dicono concordanti se a modalità elevate di uno dei due tendono ad essere associate modalità crescente, anche se tale relazione non è necessariamente lineare. Due indici di concordanza comunemente utilizzati, sono l'indice tau di Kendall e il coefficiente di Spearman.

Indici di correlazione e di concordanza tra Tasso di natalità e Tasso di mortalità[4]

Indice	Valore assunto dall'indice
Pearson Correlation	.324
Kendall's tau_b	-.039
Spearman's rho	.010

Notiamo che l'indice tau di Kendall assume addirittura un valore negativo (anche se prossimo allo zero) e che l'indice di Spearman assume un valore molto basso (e non significativamente diverso da zero). I due indici considerati sono infatti meno sensibili del coefficiente di correlazione alla presenza dei valori estremi. [*Non robustezza del coefficiente di correlazione lineare*]
Per comprendere il motivo di questi risultati contradditori, consideriamo più dettagliatamente la nuvola di punti nel diagramma di dispersione: si possono riconoscere due nuvole di punti. La prima sembrerebbe indicare una sostanziale indifferenza tra i due caratteri. La seconda nuvola ha invece un orientamento positivo, ad indicare che all'aumentare del tasso di natalità tende ad aumentare il tasso di mortalità.

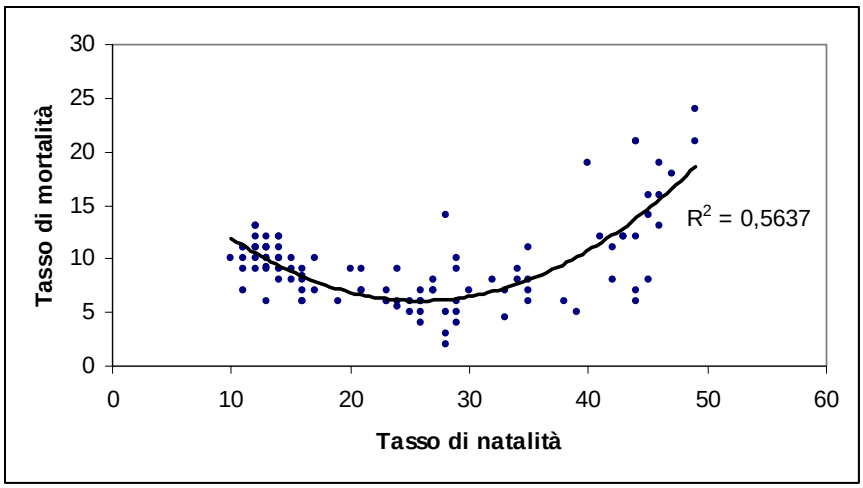
[4] Risultati ottenuti utilizzando SPSS (Analyze, Correlations, Bivariate).

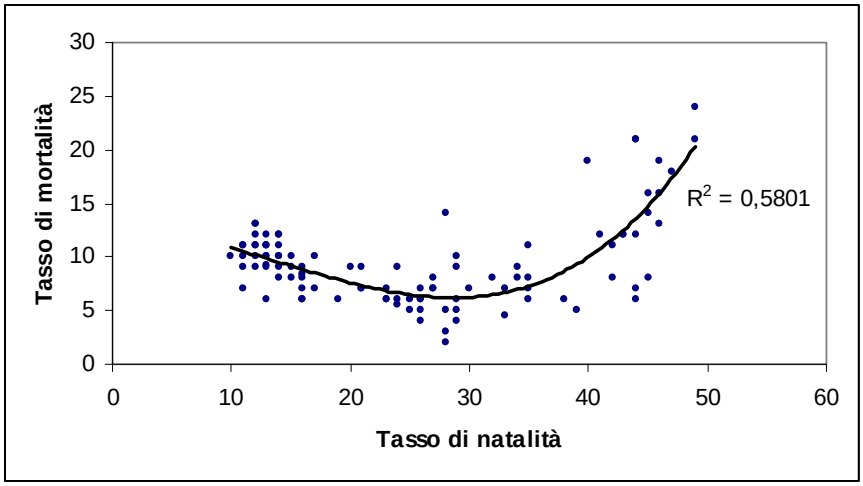


Sono in particolare i paesi dell'Africa e alcuni del Medio Oriente e della zona asiatica a far parte di questo secondo gruppo di osservazioni (in particolare, sono i paesi più poveri delle ultime due macroaree). La retta dei minimi quadrati "tenta" di adeguarsi all'intera nuvola di punti e viene attratta dai punti della seconda nuvola; il coefficiente di correlazione lineare, che misura la dispersione intorno alla retta dei minimi quadrati viene anch'esso distorto. [*Cautele nella valutazione del coefficiente di correlazione lineare*]. Le altre due misure di concordanza, meno sensibili a valori estremi, segnalano l'assenza di concordanza.

Se quindi la retta è del tutto inadeguata a sintetizzare la relazione tra i due caratteri, l'analisi del diagramma di dispersione suggerisce che un polinomio di grado superiore al primo (in particolare, una curva quadratica o cubica) descrive probabilmente meglio la relazione tra le due variabili.

A questo fine, confrontiamo modelli di regressione quadratica e di regressione cubica per spiegare il "Tasso di mortalità" in funzione del "Tasso di natalità": di seguito sono riportati i rispettivi grafici e coefficienti di determinazione. [*Interpolazione di una nuvola di punti: quale funzione scegliere?*]





Si osserva un buon adattamento ai dati sia della curva quadratica che di quella cubica; il calcolo dei corrispondenti coefficienti di determinazione, pari a 0.564 per la prima e a 0.580 per la seconda, indica che è leggermente preferibile quest'ultima (si ricordi che il coefficiente di determinazione relativo alla regressione lineare è pari a 0.105).
[*L'impatto della variabile esplicativa: caso lineare e non lineare*]

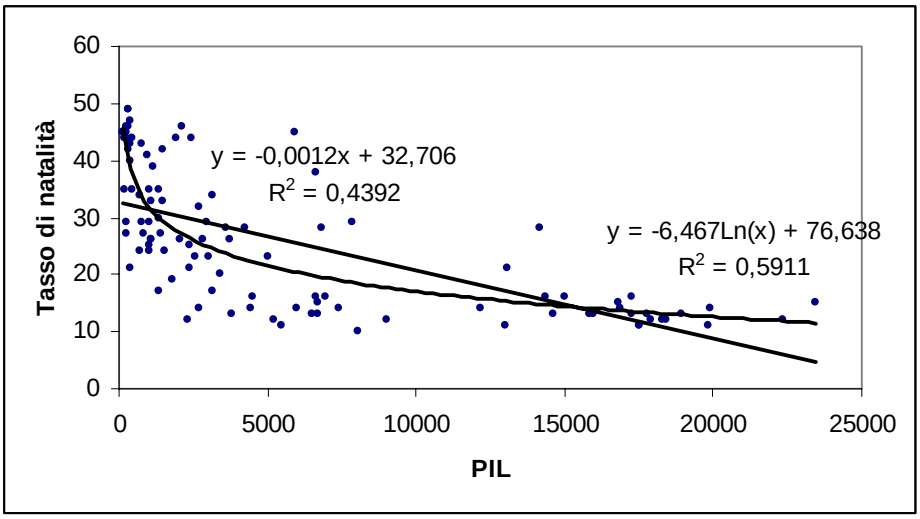
Sostanzialmente, paesi con alti tassi di natalità hanno usualmente anche un alto tasso di mortalità, così come paesi con i più bassi tassi di natalità mostrano comunque livelli di mortalità abbastanza elevati. Se pensiamo al tasso di mortalità come indicatore del benessere sanitario di un paese ci aspetteremmo valori bassi dell'indice per i paesi industrializzati, e valori via via più alti per paesi in via di sviluppo. Allo stesso tempo, ci aspettiamo che siano proprio i paesi poveri quelli in cui si mettono al mondo più figli. La relazione qui osservata, quindi, ci sorprende un poco (mentre sarebbe stata più naturale una relazione lineare).
Per comprendere meglio questa relazione osservata (e quindi identificare nel grafico dove si collocano i paesi ricchi e i paesi poveri) valutiamo l'eventuale associazione di ciascuna di queste due variabili con il PIL.
Il tasso di natalità, come si evince dal coefficiente di correlazione lineare corrispondente, pari a -0.663, ha una notevole associazione lineare (inversa) col PIL. La concordanza tra i due caratteri è confermata anche dall'indice tau di Kendall (concordanza media) e dal coefficiente di Spearman[5].

Indici di associazione (concordanza) tra PIL e Tasso di natalità e tra PIL e Tasso di mortalità

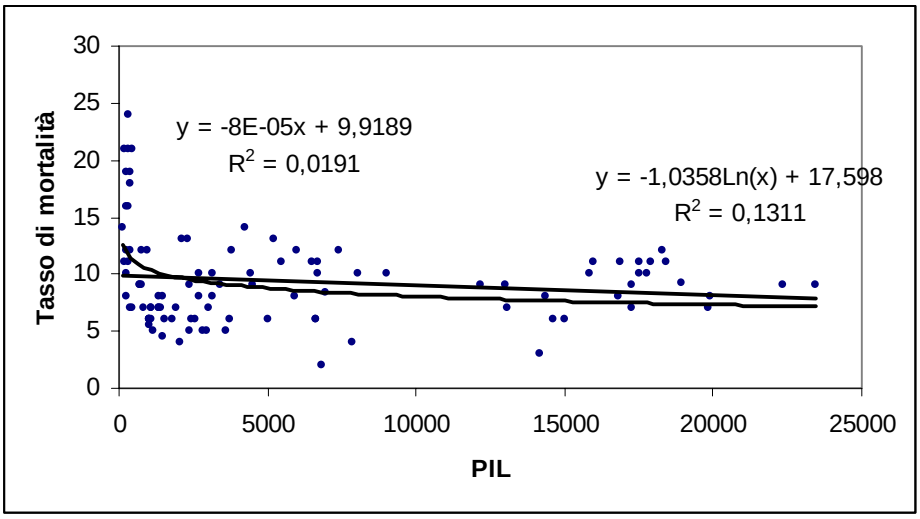
Indice	PIL, Tasso natalità	PIL, Tasso mortalità
Pearson Correlation	-0,66	-0,14
Kendall's tau_b	-.583	-.097
Spearman's rho	-.776	-.150

I grafici di dispersione del tasso di natalità e, rispettivamente, del tasso di mortalità sul PIL confermano questa indicazione. Il primo dei due mostra come la retta dei minimi quadrati si adatti discretamente ai dati; anche in questo caso, naturalmente, si potrebbero cercare altre curve per migliorare l'adattamento. Ad esempio, la curva logaritmica mostra un migliore adattamento (nel grafico sono riportate la retta di regressione e la curva logaritmica).

[5] Risultati ottenuti utilizzando SPSS (Analyze, Correlations, Bivariate).



Passando ora a considerare PIL e tasso di mortalità, notiamo dal grafico che l'associazione tra i due caratteri è piuttosto debole. Il valore così basso del coefficiente di correlazione è quindi affidabile e riflette una debole associazione lineare.
[*Cautele nella valutazione del coefficiente di correlazione lineare e di determinazione*]



Quindi, sia il tasso di natalità che il tasso di mortalità diminuiscono al crescere del PIL; in particolare, c'è una discreta associazione lineare (inversa) tra tasso di natalità e PIL.
Naturalmente, la presenza di associazione lineare di grado abbastanza elevato tra tasso di natalità e PIL non implica necessariamente la presenza di nesso di causalità tra le due variabili; ovvero, non si può affermare che PIL bassi *provocano* alti tassi di natalità e viceversa. [*Associazione e causalità*]

Quello che però abbiamo notato da questi ultimi due grafici è che i paesi sviluppati, oltre ad avere i più bassi tassi di natalità, hanno anche tassi medi di mortalità, il che, almeno a prima vista, sorprende. In realtà, è possibile pensare che popolazioni dove il numero di nati è esiguo (ricordiamo che il tasso di natalità non è altro che numero di nati registrati per ogni mille abitanti) le età più anziane pesano di più, e quindi pesano di più quelle fasce di popolazione dove la mortalità è più diffusa. In altri termini, il tasso di mortalità (numero di morti per mille abitanti) è più elevato non perché si muore "più intensamente", ma perché si è "più vecchi".

4. *Come cresce (o decresce) la popolazione? Come variano gli indicatori di incremento demografico nelle diverse macroaree?*

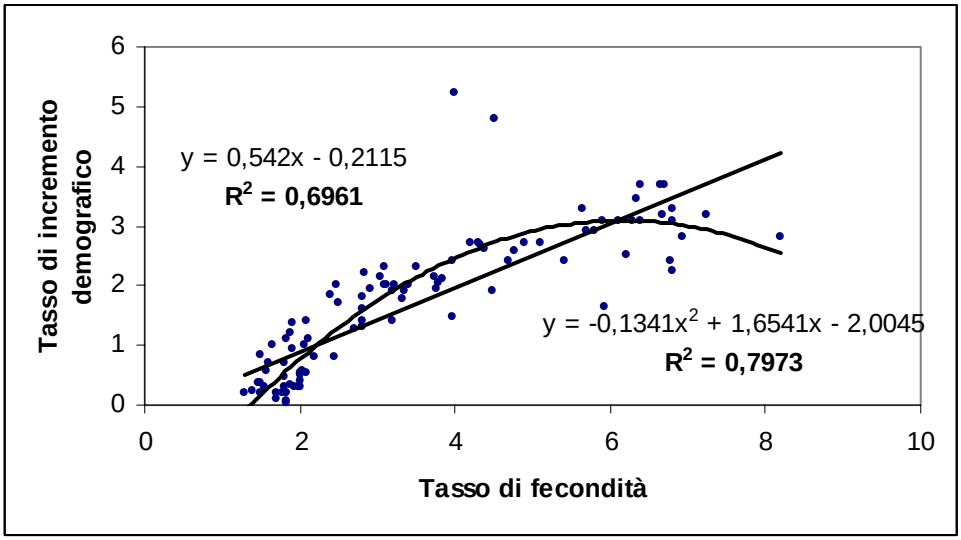
Per analizzare l'incremento della popolazione, nell'intero collettivo o nelle singole macroaree, consideriamo ovviamente la variabile "tasso di crescita" (espresso come aumento annuo percentuale della popolazione), e ne calcoliamo indicatori di posizione: la tabella che segue mostra questi indicatori sia con riferimento all'intero collettivo che nelle 6 macroaree considerate.
[*Analisi stratificata: le misure di sintesi*]

Case Summaries

Tasso di incremento demografico (annuo)

Macroarea (regione o	N	Mean	Median	Std. Deviation
OECD	21	,511	,400	,338
East Europe	12	,204	,250	,320
Pacific/Asia	14	1,539	1,690	,804
Africa	18	2,719	2,805	,468
Middle East	16	2,914	2,800	1,151
Latn America	21	1,877	2,000	,676
Total	102	1,664	1,790	1,195

Si nota, dalla tabella, che il tasso di crescita annuo medio, sull'intero collettivo, è 1.664 persone su 100e che questo però si differenzia notevolmente tra le varie macro aree, essendo notevolmente più basso nei paesi dell'Europa orientale e dell'OECD e significativamente più alto nei paesi dell'Africa e del Medio Oriente.
Ovviamente, anche il tasso di fecondità può essere un importante indicatore concernente l'incremento demografico. Per valutare se e come questa variabile è associata al tasso d'incremento demografico, consideriamo il grafico di dispersione del tasso d'incremento demografico su quello di fecondità:

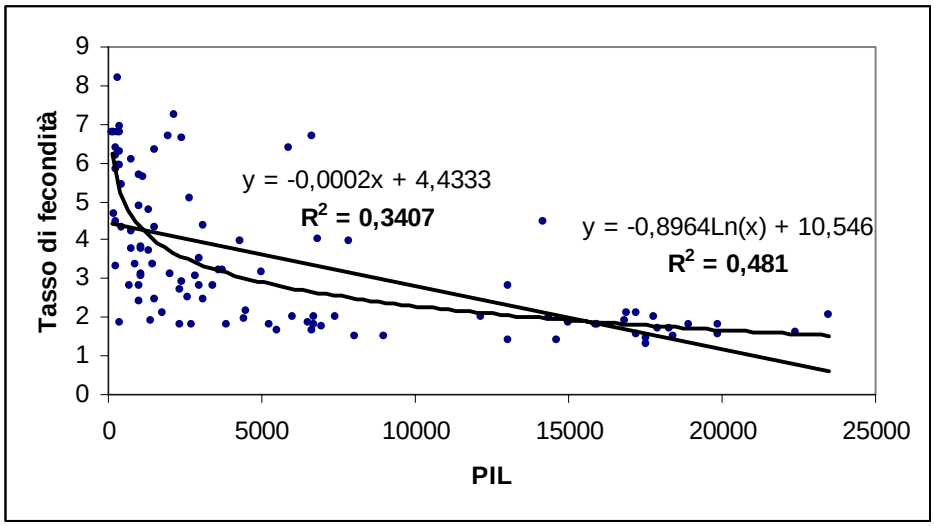


E' evidente una forte associazione lineare diretta tra le due variabili, confermata dal valore del coefficiente di correlazione lineare, pari a 0.834; la retta dei minimi quadrati descrive quindi piuttosto bene la dipendenza tra il tasso d'incremento demografico ed il tasso di fecondità. È anche immediato osservare come siano presenti due Paesi che appaiono come outliers rispetto alle variabili prese in esame; dal grafico si deduce che si tratta di Paesi con tasso di fecondità piuttosto basso (intorno al 4%) e tasso di incremento demografico molto elevato. E' facile rilevare (attraverso le opzioni del grafico, o direttamente dall'analisi del dataset, che si tratta degli Emirati Arabi Uniti e

del Kuwait per i quali non la fecondità ma l'immigrazione ha effettivamente costituito uno dei motori più forti dell'incremento demografico**.**

Un'altra considerazione interessante è la seguente: nel diagramma di dispersione è riportata anche una curva polinomiale di grado 2, che risulta più efficace della retta nella spiegazione del Tasso di incremento demografico in funzione del Tasso di fecondità. [*Interpolazione di una nuvola di punti: quale funzione scegliere?*] Tale curva evidenzia che, a partire da un tasso di fecondità pari circa a 6, la relazione "diretta" tra i due caratteri si attenua: un aumento del tasso di fecondità si traduce in un aumento del tasso di incremento demografico inferiore rispetto a quanto non accada se si considerano tassi di fecondità più bassi. Questo probabilmente è dovuto a tassi di mortalità più elevati nei paesi in via di sviluppo.
[*L'impatto della variabile esplicativa: caso lineare e non lineare*]

Interessante può anche essere valutare se, quanto ed in che modo il tasso di fecondità è legato ad un indicatore generale di ricchezza quale è il PIL; essendo il coefficiente di correlazione lineare pari a –0.584, si rileva una discreta associazione lineare inversa tra i due caratteri. In realtà, anche in questo caso una regressione logaritmica appare preferibile per spiegare la variabile dipendente, come si può evincere dall'analisi del grafico di dispersione riportato di seguito con le due curve (lineare e logaritmica) e i rispettivi coefficienti di determinazione.



Dal grafico si nota comunque la presenza di alcune osservazioni "eccezionali", che tendono a "spostare" la curva verso l'alto.
In ogni caso, la relazione inversa rilevata indica che, tendenzialmente, Paesi più poveri hanno tassi di fecondità più elevati. Del resto, anche una semplice analisi stratificata del tasso di fecondità, con la rilevazione delle medie e delle mediane per macroarea, fornisce indicazioni in questo senso.
[*Analisi stratificata: le misure di sintesi*]

Case Summaries

Tasso di fertilità (numero medio di figli per donna)

Macroarea (regione o	N	Mean	Median
OECD	21	1,746	1,800
East Europe	12	1,893	1,855
Pacific/Asia	14	2,914	2,600
Africa	18	6,047	6,245
Middle East	16	4,611	3,980
Latn America	21	3,336	3,080
Total	102	3,460	2,865



vicino allo 0. L'analisi del grafico di dispersione consente in questo caso di confermare che i due caratteri non sono associati linearmente. Anche se un coefficiente di correlazione prossimo allo zero e una retta di regressione con pendenza quasi nulla non consentono di concludere che tra i due caratteri non esista un altro tipo di associazione – non lineare – in questo caso si osserva che i due caratteri non sembrano associati.

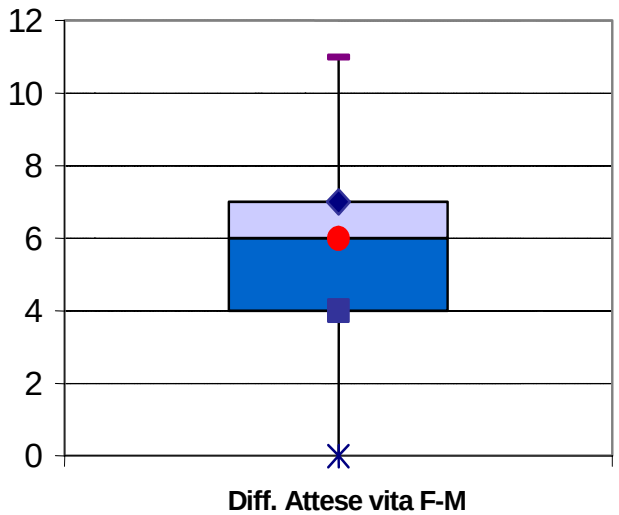
Sembra doveroso a questo punto sottolineare che tutte le considerazioni fatte in merito al tasso di istruzione superiore vanno trattate con una certa cautela. Infatti, il carattere scelto per misurare il fenomeno di interesse è il numero di studenti universitari ogni 1000 abitanti. In sostanza, è una percentuale valutata sulla *totalità della popolazione*. Tale percentuale dipenderà necessariamente *anche* dalla struttura per età della popolazione. In particolare, se in una popolazione sono le fasce di età più elevate quelle prevalenti, un tasso di istruzione universitaria molto basso sarebbe "inevitabile" (gli studenti universitari sono pochi perché sono pochi i "giovani"). Un indicatore più sensato sarebbe probabilmente dato dal numero di studenti universitari sul numero totale di soggetti che potrebbero almeno virtualmente iscriversi all'università.

6. *Cosa si può dire, in generale, sulla differenza tra condizione maschile e condizione femminile rispetto ai diversi aspetti considerati nell'analisi?*

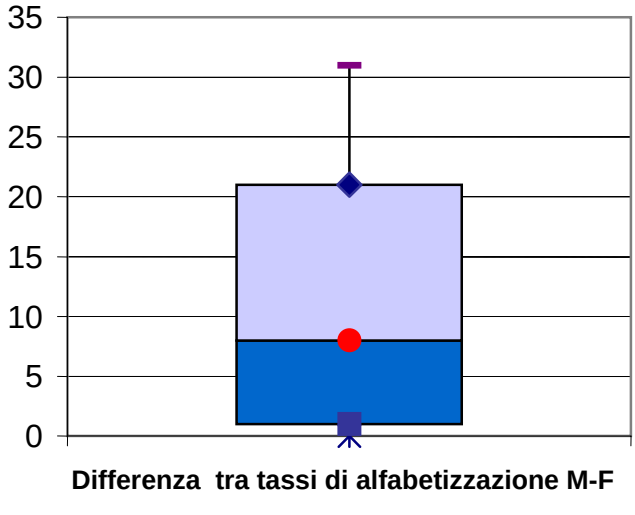
E' interessante chiedersi se la differenza tra condizione maschile e femminile, rispetto a diversi aspetti, è legata ad indicatori complessivi di benessere. A questo scopo, definiamo due variabili come differenze tra attese di vita (attesa di vita *femminile* – attesa di vita *maschile*) e tassi di alfabetizzazione (tasso di alfabetizzazione *maschile* – tasso di alfabetizzazione *femminile*). Indicatori sintetici di queste due variabili sono forniti nel seguito[6]:

		Differenze tra tassi di alfabetizzazione M-F	Differenze tra attese di vita F-M
N	Validi	82.0000	102.0000
	Mancanti	20.0000	0.0000
Media		11.0122	5.3529
Mediana		8.0000	6.0000
Varianza		104.9998	4.6465
Deviazione standard		10.2469	2.1556
Campo di variazione		31.0000	11.0000
Quartili	Primo	1.0000	4.0000
	Terzo	21.0000	7.0000

Come già osservato in precedenza, rispetto a questi due aspetti nessuno degli stati considerati presenta condizioni migliori per le femmine (il minimo è, per entrambe le variabili, pari a 0). Si osserva inoltre una variabilità nettamente maggiore per la differenza tra i tassi di alfabetizzazione che presenta inoltre, a differenza dell'altra variabile, una certa asimmetria verso destra. I due box plot[7] confermano queste indicazioni (il primo è relativo alla differenza tra le attese di vita, il secondo alle differenze tra i tassi di alfabetizzazione).
[*Box plot: una rappresentazione sintetica della distribuzione*]



Diff. Attese vita F-M

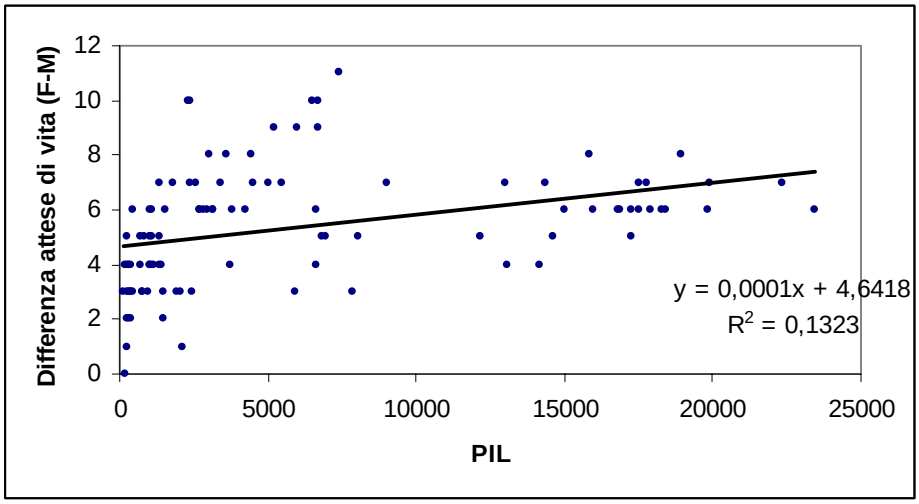


Differenza tra tassi di alfabetizzazione M-F

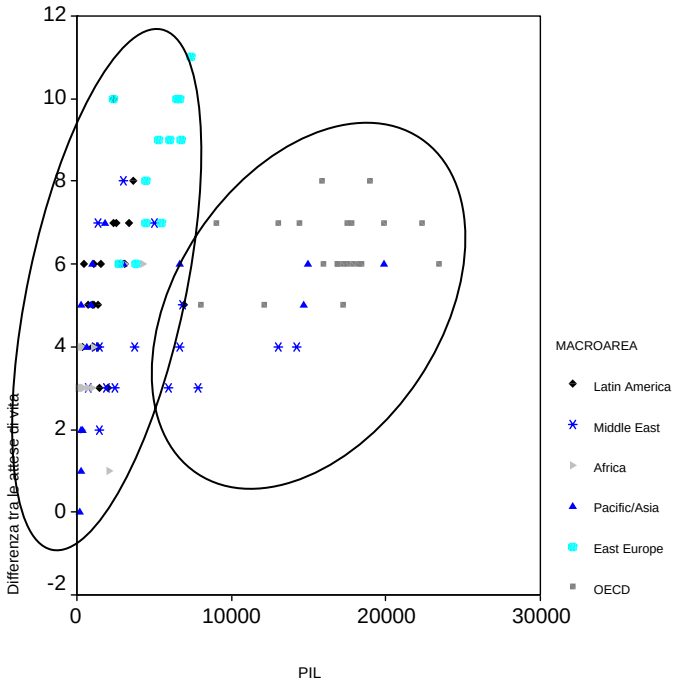
Bisogna di nuovo prestare attenzione al fatto che per questo secondo box plot si hanno valori mancanti in corrispondenza dei paesi dell'OECD per i quali, come già osservato in precedenza, si potrebbe ipotizzare una differenza nulla o piuttosto bassa tra tassi di alfabetizzazione femminile e maschile. In sostanza, quindi, l'aspetto del box plot potrebbe anche dipendere dal fatto che tali valori mancanti "impoveriscano" la coda di sinistra (nella quale si collocherebbero se fossero stati osservati)**.** [*I valori mancanti*]

[6] Risultati ottenuti con SPSS (Analyze, Descriptive statistics, Frequencies).
[7] Risultati ottenuti con EXCEL (macro Stat4038).

Consideriamo ora il grafico di dispersione della differenza delle attese di vita sul PIL; questo indica effettivamente una certa associazione positiva tra le due variabili.

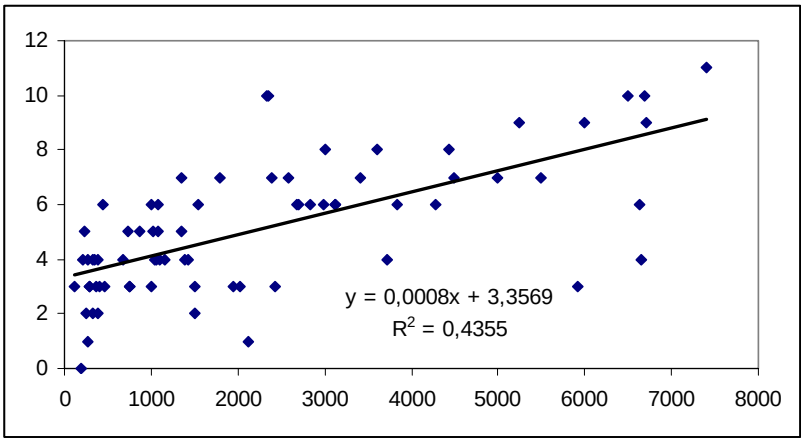


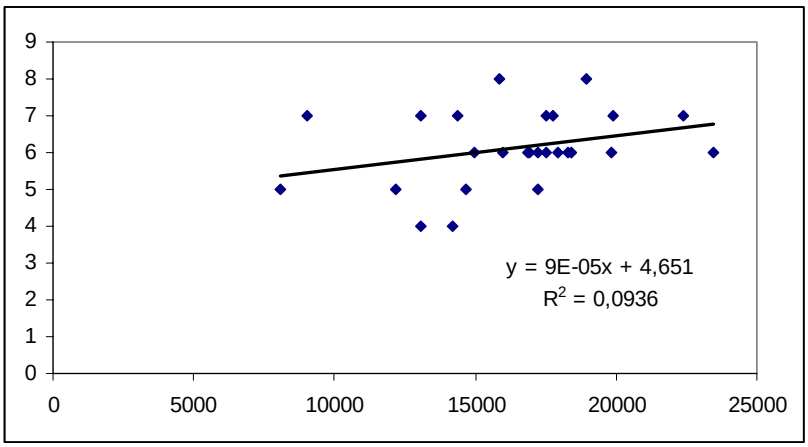
Osservando più attentamente il diagramma di dispersione, possiamo notare che la nuvola dei punti individuati dalle modalità dei due caratteri è sostanzialmente costituita da 2 nuvole. Una prima nuvola di punti è relativa ai paesi con PIL basso, per i quali si osserva una relazione positiva tra la differenza fra le attese di vita ed il PIL. Per il secondo gruppo, quello costituito da paesi con PIL elevato, si osserva invece una sostanziale assenza d'associazione lineare. Questo può essere spiegato molto semplicemente con il fatto che per paesi che hanno comunque un livello di ricchezza elevato, la differenza tra le attese di vita si attesta intorno ad un valore che non varia molto al variare del PIL.



Notiamo che le due nuvole separano i paesi in base alla macro area di appartenenza e al livello del PIL pro-capite. La nuvola con inclinazione più marcata è quella relativa ai paesi più poveri (che appartengono alle aree più "critiche"). Notiamo invece che la nuvola con inclinazione meno marcata è quella relativa ai paesi delle macroaree più "forti" e ai paesi più ricchi delle macroaree

più “deboli” (ad esempio, i paesi del medio oriente che appartengono a questo secondo gruppo sono Israele ed Emirati Arabi).



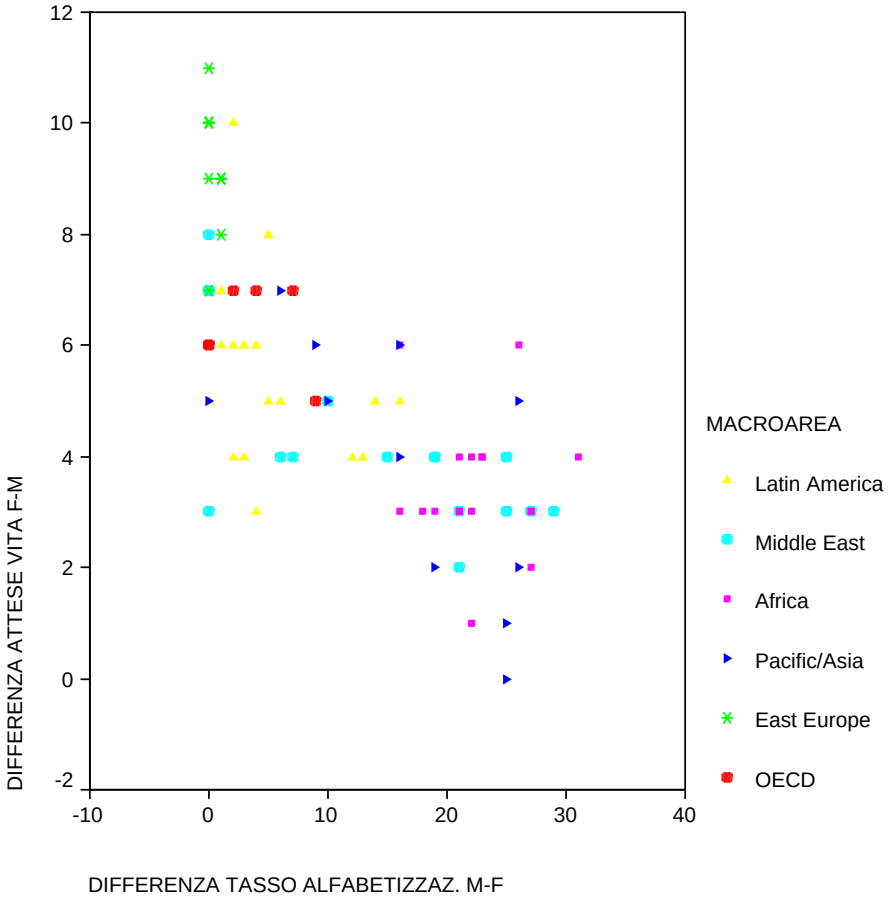


Di fianco sono riportati i diagrammi di dispersione relativi ai due gruppi.
Questi due grafici confermano la sensazione iniziale: nei paesi con basso PIL pro-capite una variazione del PIL porta ad una variazione abbastanza marcata della differenza tra le attese di vita. Nei paesi con PIL pro-capite elevato, questa “reazione” è molto più tenue (si noti che la capacità esplicativa della retta di regressione in questo secondo caso è decisamente bassa: il coefficiente di determinazione R^2= 0.09).

Di seguito riportiamo il coefficiente di correlazione lineare e gli indici di concordanza, l’indice tau di Kendall e il coefficiente di Spearman tra il PIL e le due differenze considerate. Con riferimento alla coppia PIL, differenza tra attese di vita, il coefficiente di correlazione risulta pari a 0.364, valore non è molto significativo, dal momento che risulta dall’aggregazione di strutture diverse di associazione.

		PIL	Diff. attese vita F-M	Diff. tasso di alfabetizz. M-F
Pearson Correlation	PIL	1	.364	-.340
	Diff. attese vita F-M	.364	1	-.703
	Diff. tasso di alfabetizz. M-F	-.340	-.703	1
Kendall's tau_b	PIL	1	.415	-.351
	Diff. attese vita F-M	.415	1	-.566
	Diff. tasso di alfabetizz. M-F	-.351	-.566	1
Spearman's rho	PIL	1.000	.577	-.487
	Diff. attese vita F-M	.577	1.000	-.711
	Diff. tasso di alfabetizz. M-F	-.487	-.711	1.000

E’ anche interessante notare come sia elevato, e di segno negativo, il coefficiente di correlazione lineare tra la differenza tra le attese di vita e la differenza tra i tassi di alfabetizzazione; in realtà, l’analisi stratificata, per macroarea, riportata nel grafico di seguito, mostra come vi sia disomogeneità tra le macroaree stesse, con riferimento ai caratteri in esame.


12
10
8
6
4
2
0
-2
-10
0
10
20
30
40
DIFFERENZA ATTESE VITA F-M
DIFFERENZA TASSO ALFABETIZZAZ. M-F
MACROAREA
Latin America
Middle East
Africa
Pacific/Asia
East Europe
OECD