Ing. Fisica
Politecnico di Milano
a.a. 2020-2021
Corso di Statistica - Prof. Toigo

Non fidarsi è bene, fidarsi è meglio

Analisi statistica sul livello di fiducia riposto nei risultati scientifici

Cecchettini Stefano,

Martellini Sara,

Rossi Mercanti Davide

Introduzione

Obiettivo di questa analisi statistica è studiare come varia l'indice di fiducia della popolazione nei confronti dei risultati scientifici, con un approccio che parte da una prospettiva mondiale, fino a concentrarci nel dettaglio sulla situazione in Italia. In particolare, cercheremo di trovare un modello per la deduzione del grado di fiducia di un individuo in funzione dei dati personali del soggetto e capire quali di essi siano davvero influenti.

Il nostro dataset proviene dal UK Data Service ed è il risultato della ricerca *"Wellcome Global Monitor",* realizzata nel 2018 dal *"Wellcome Trust",* ente di beneficienza inglese, e "*The Gallup Organization Ldt"*, società americana di analisi.

La ricerca è stata condotta tramite un questionario di 30 domande volto a misurare il livello di conoscenza e fiducia nella scienza oltre alla visione dei benefici (personali e sociali) che questa può offrire. Per ogni soggetto esaminato, inoltre, sono stati registrati altri dati riguardanti il suo profilo: età, sesso, grado di educazione, fascia di reddito, tipo di lavoro, area in cui risiede e Paese di provenienza.

Inizieremo quindi con un'analisi secondo la statistica descrittiva sul grado di fiducia definito, dividendo il nostro dataset per macroaree geografiche. Effettueremo poi dei test di ipotesi volti a dimostrare la veridicità o meno di nostre assunzioni sull'indice in questione, concentrandoci anche su fette della popolazione divise secondo criteri differenti. Infine, nel capitolo della regressione e grazie ai risultati precedenti, valuteremo anche altri indici, per determinare quanto essi siano correlati al grado di confidenza nella scienza dell'italiano medio.

Questo progetto è stato realizzato con l'ausilio del *software statistico* ***R***.

Indice

1-Dataset

1.1 Presentazione

Il nostro dataset comprende le risposte ad ogni domanda del questionario e le informazioni aggiuntive sul soggetto. Sulla base delle risposte ad ogni categoria di domande, è stato quindi assegnato ad ogni intervistato un valore corrispondente di ***WGM_Index: Confidence level*** nella scienza. Il numero di stati presenti è 144, per un totale di circa 140 mila risposte.

La nostra analisi tuttavia si concentrerà su due dataset in particolare, relativi a due macroaree geografiche: **l'Europa** e il **Medio Oriente**. Questa scelta è dovuta al voler prendere in esame due regioni che si distinguessero dal punto di vista culturale e politico, ma allo stesso tempo vicine geograficamente e legate da importanti rapporti commerciali ed economici. In questo modo contiamo di ottenere risultanti interessanti da inserire in un contesto di attualità.

Il nostro dataset si riduce a 53 stati, con un totale di poco più di 42 mila soggetti intervistati.

1.2 Definizione di indice di fiducia

In alcune domande del questionario viene richiesto ai soggetti di esprimere il loro grado di interesse, conoscenza o scetticismo su varie questioni, selezionando una tra:

□ A lot □Some □Not much □Not at all □Don't know □Refused

Da una valutazione pesata delle risposte, è stato quindi estrapolato un valore numerico decimale compreso tra 1 e 4, volto a valutare il ***Confidence level*** nella scienza nel suo complesso.
E' interessante fare una piccola precisazione sui termini scienza e scienziato utilizzati in questo contesto, così da cogliere le sottili sfumature fra indici diversi: per *scienza* si intende la *comprensione che abbiamo del mondo* derivata da *osservazioni e verifiche*; per *scienziati* ci si riferisce alle persone che studiano *il pianeta Terra, la natura e la medicina* fra le altre cose.
Seguono alcune delle domande più degne di nota come esempio delle categorie affrontate.

- Hai provato personalmente ad informarti su argomenti scientifici negli ultimi 30 giorni?
- Ti aspetti che gli scienziati trovino informazioni accurate sul mondo?
- Pensi che gli scienziati in questo Paese facciano il loro lavoro con l'intento di beneficiare la gente comune?
- Ti fidi dei consigli medici e di salute da assistenti medici come dottori e infermieri, in questo Paese?
- Quanto pensi che gli scienziati siano onesti su chi stia pagando per il loro lavoro?
- In generale pensi che il lavoro degli scienziati benefici molte, alcune o poche persone in questo Paese?
- Complessivamente, pensi che la scienza e la tecnologia aiutino a migliorare la vita per le prossime generazioni?
- Cosa pensi della seguente affermazione: i vaccini sono efficaci. Sei fortemente o in parte d'accordo o in disaccordo?

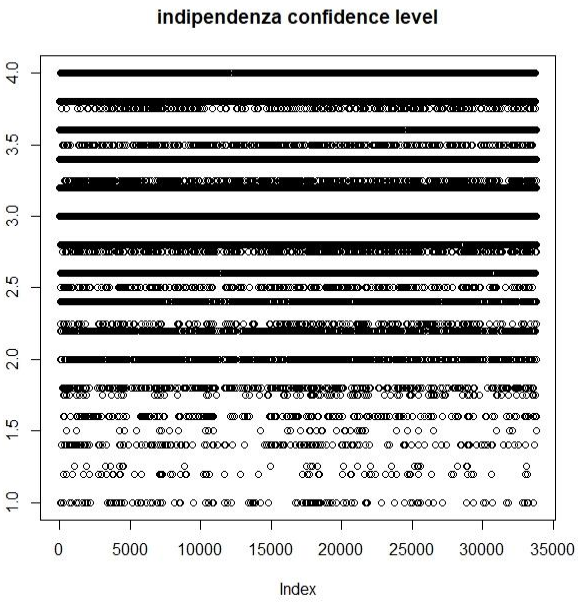
Degli altri indici che abbiamo deciso essere rilevanti per la nostra indagine, riportiamo la corrispondenza fra dato numerico presente nel nostro dataset e l'etichetta, cioè la casella associata contrassegnata dal soggetto. Questi indici assumono tutti un valore crescente e proporzionale alla corrispettiva etichetta (tranne per *gender* che è binario, per *employment status* e per *area type*) e sono:

- **View of Scientists:** indice della fiducia soltanto negli **scienziati**
    1. Low trust
    2. Medium trust
    3. High trust
- **View of Science:** come il soggetto vede soltanto i **benefici** della scienza per la persona e per la società
    1. Enthusiast
    2. Included
    3. Excluded
    4. Sceptic
- **Age**
- **Age Categories:**
    1. Under30: 15-29 anni
    2. Second adult age: 30-49 anni
    3. Over 50: 50+ anni
- **Gender:**
    1. Male
    2. Female
- **General Educational Background:**
    1. Elementary education: 0-8 anni di educazione
    2. Secondary education: 9-15 anni
    3. University or college level education: 15 + anni
- **Per Capita Income Quintiles (Household Income)**: il quintile di appartenenza a seconda della fascia di reddito
    1. Poorest 20%
    2. Second 20%
    3. Middle 20%
    4. Fourth 20%
    5. Top 20%
- **Employment Status:**
    1. Employed full time for an employer
    2. Employed full time for self
    3. Employed part time do not want full time
    4. Employed part time want full time
    5. Unemployed
    6. Out of workforce
- **Area Type:**
    1. Lives in rural area or small town
    2. Lives in city or suburb of city

1.3 Ipotesi sul dataset

Possiamo ragionevolmente assumere che il campione dell'Indice di fiducia relativo ai vari stati è formato da variabili aleatorie indipendenti tra loro e identicamente distribuite. Vale a dire che le risposte date da un soggetto non influenzano quelle di un altro, così come le rilevazioni in stati diversi.

L'ipotesi d'indipendenza può essere verificata dall'assenza di *trend* particolari nello *scatterplot* dei campioni. Riportiamo ad esempio lo *scatterplot* del *Confidence Level* effettuato sull'Europa. Nonostante l'aspetto "a righe", imputato al fatto che i valori sono discreti e non assumono tutte le cifre decimali, l'omoschedasticità delle rilevazioni è comunque verificata. Infatti, i valori non sembrano disporsi secondo alcun *pattern* ascendente o discendente.



Inoltre, trattandosi di un questionario anonimo è ragionevole dare per assunto che le rilevazioni siano state prese nella maniera più professionale possibile, senza nessuna influenza tra soggetti diversi e lasciando loro liberi di rispondere in maniera sincera. Possiamo quindi continuare l'analisi statistica, considerando il nostro dataset composto da variabili indipendenti e identicamente distribuite.

2-Statistica descrittiva

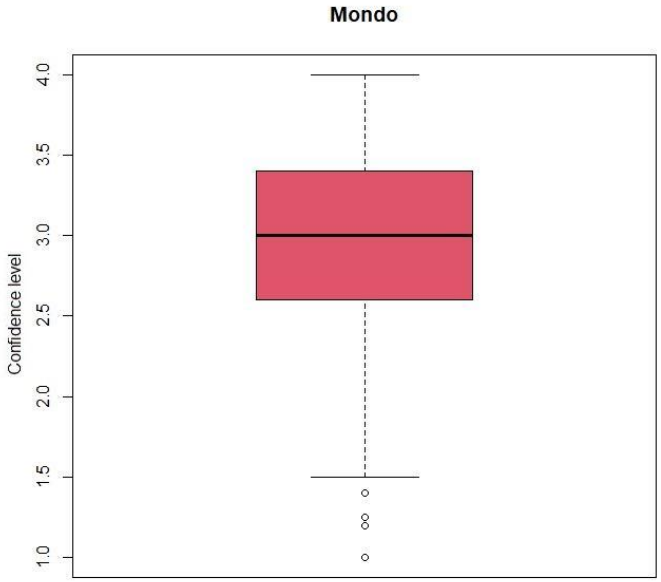
In questa sezione analizzeremo nel dettaglio il nostro dataset, con lo scopo di evidenziarne le caratteristiche di base e ricavarne una sintesi semplice del campione e delle misure raccolte.

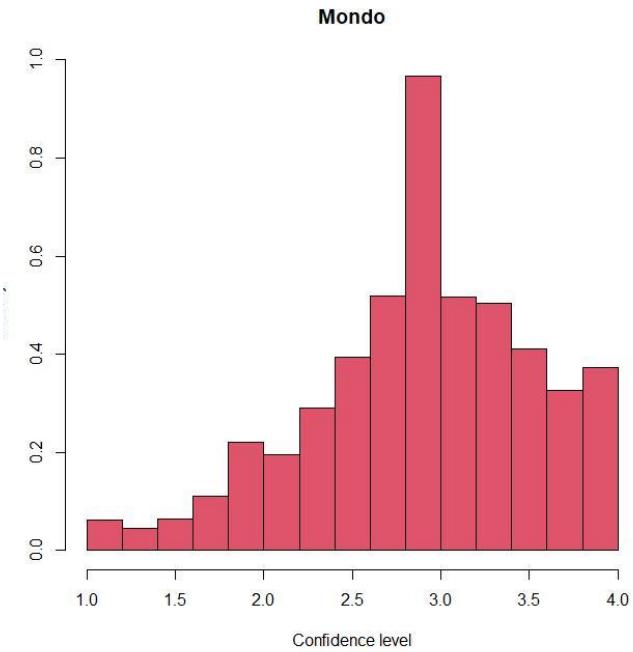
2.1 Indice di fiducia

Essendo il nostro primario interesse, se ne vuole analizzare anzitutto la distribuzione ***globale***:

```
> summary(mondo$WGM_Index)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   2.600   3.000   2.983   3.400   4.000
```

Con una media, anche se di poco, inferiore alla mediana, viene segnalata la presenza di una coda sinistra, confermata dall'*istogramma*. L'*IQR* risulta molto ampio, indice del grande numero di dati preso in considerazione. La mediana, il cui valore è uguale a 3.0, risulterà la stessa in tutti i continenti, come si riscontra nei grafici.

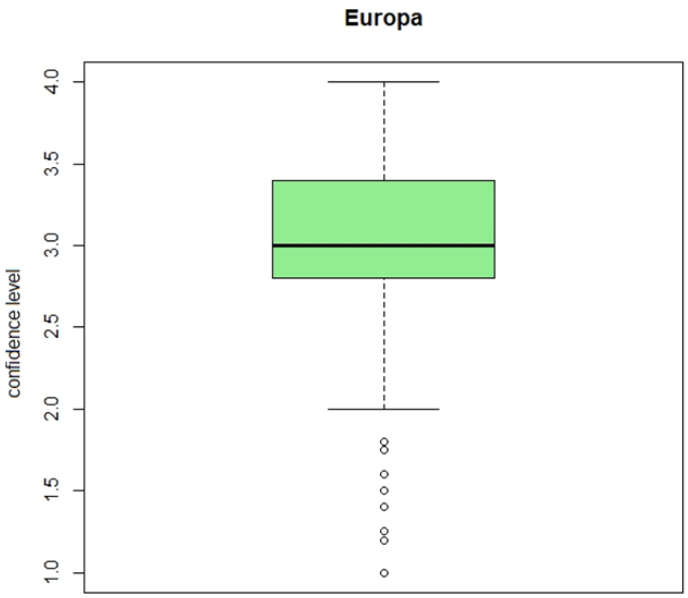


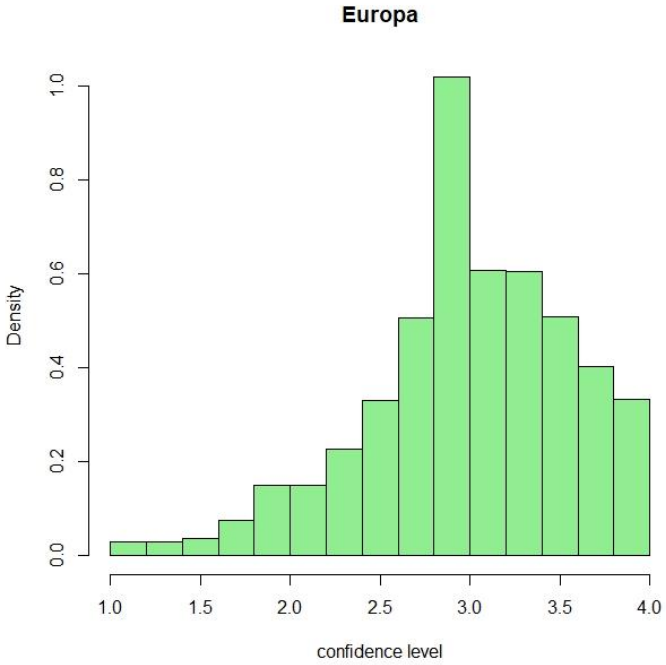


La nostra analisi prosegue nel dettaglio con i due continenti selezionati.
Seguono i risultati ottenuti sull*'Europa*, dataset formato da 40 stati, ognuno di 900 campioni circa.

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   2.800   3.000   3.083   3.400   4.000
```

Con una media di 3.083, superiore a quella mondiale, risulta uno tra i continenti con maggior fiducia. Nonostante la mediana sia inferiore, dal grafico non si rileva alcuna coda destra: avendo una grande quantità di dati presi in considerazione ed essendo la differenza fra i due valori estremamente bassa, la prova grafica non risulta più efficiente. L'*IQR* risulta meno ampio del rispettivo globale, altro indice di una maggior fiducia da parte degli europei. Si conclude lo stesso anche dal *boxplot*, ben diverso dal medesimo mondiale per la presenza di un baffo sinistro più contenuto, nonostante il numero di *outliers* sia leggermente aumentato.

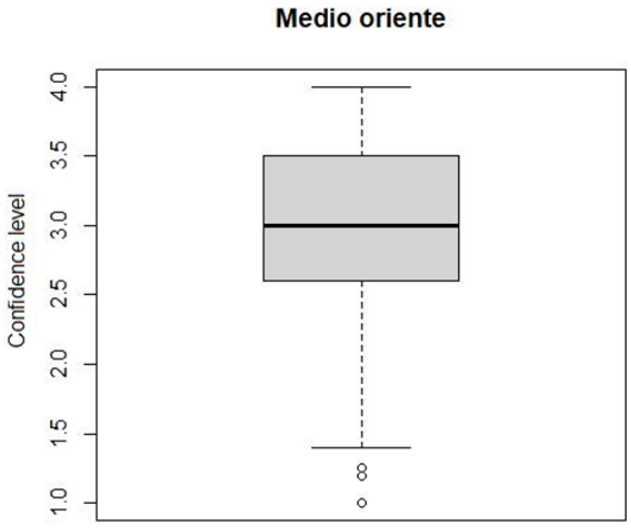


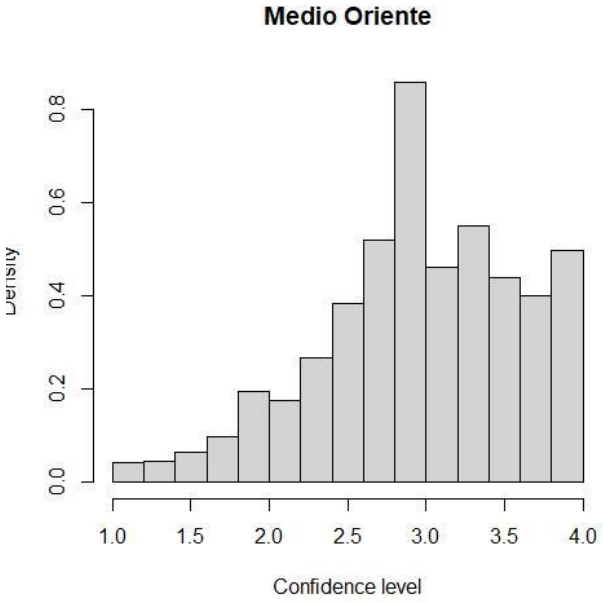


Si vogliono ora commentare secondo gli stessi criteri i dati relativi al *Medio Oriente*:

```
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 1.000   2.600   3.000   3.044   3.500   4.000
```

La media di 3.044 è lievemente inferiore rispetto a quella europea: se ne può dedurre che generalmente il Medio Oriente è meno fiducioso nella scienza rispetto al nostro continente, ma con dei valori pur sempre migliori rispetto alla popolazione mondiale. L'*IQR* è leggermente superiore a quello del resto del mondo, con il terzo quantile maggiore di un decimo. Dal *boxplot* si osserva inoltre una notevole asimmetria dei baffi, con una pronunciata coda sinistra. La dispersione dei dati è probabilmente indice della diversità fra i vari stati che formano questa macroarea.

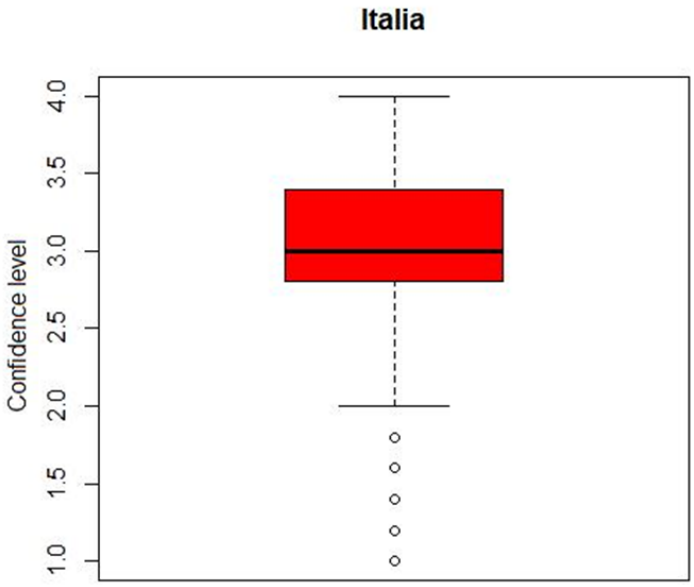


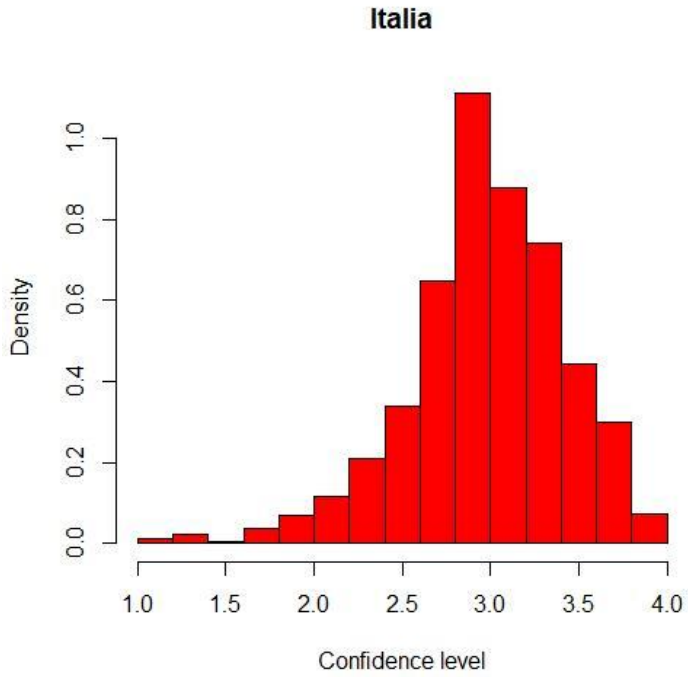


Ci concentriamo infine sui dati italiani (circa 900 campioni analizzati) per poterli mettere a confronto con i risultati precedentemente ottenuti.

```
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 1.000   2.800   3.000   3.076   3.400   4.000
```

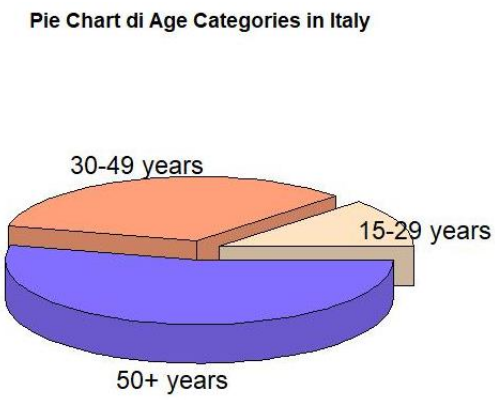
Con un valore di 3.076, la media dell'*Italia* risulta inferiore a quella europea. Dall'*istogramma* si rileva una coda sinistra meno accentuata, al contrario di quella destra ben più visibile dei grafici precedenti. Nonostante l'inferiore numero di dati considerati, l'*IQR* è uguale a quello europeo e, come mostra il *boxplot*, i baffi sono perlopiù uguali. Ciò è dovuto ad una distribuzione più concentrata, influenzata dal minor numero di campioni presi in considerazione.

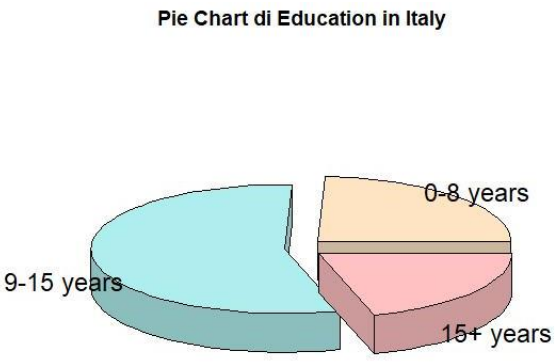




2.2 Altri indici: età, educazione e reddito

Oltre al *Confidence Level* è interessante studiare la distribuzione di altri indici fra quelli presenti nel dataset: in particolare analizziamo l'età dei soggetti intervistati, il loro grado di educazione e il reddito pro-capite dichiarato. Qui di seguito riportiamo e commentiamo solamente i valori italiani. Gli altri grafici dei continenti vengono invece mostrati a pag.22.



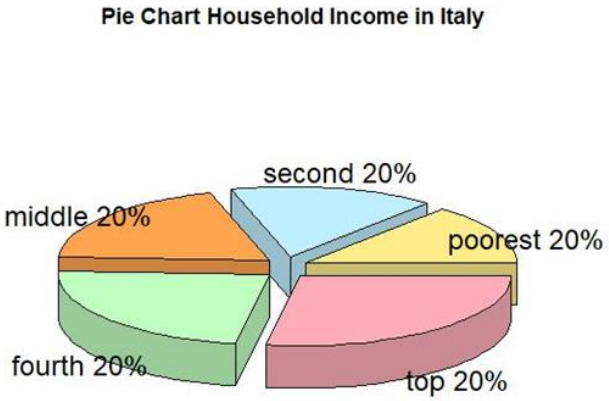


Il *pie chart* a sinistra mostra come sia stata varia, in termini di età, la partecipazione al sondaggio. Le percentuali ricalcano approssimativamente quelle di tutta la popolazione italiana, permettendoci di estendere le nostre conclusioni ad un'intera ottica nazionale. [1*]

Dal secondo grafico emerge un livello di istruzione considerevole: i campioni si concentrano sulla fascia media, con più del 75% dei partecipanti con almeno 9 anni di istruzione (primo anno di scuola secondaria superiore) e poco meno del 20% di laureati.

Commentiamo infine i risultati ottenuti sul reddito. Al contrario dell'educazione che si concentra sulla fascia media, esso risulta egualmente distribuito sulla popolazione presa in considerazione. L'equipartizione in base al *Household Income* è indice dell'attenzione posta dagli esaminatori nella scelta dei partecipanti: tutte le fasce di reddito sono rappresentate a pieno.

Questi aspetti verranno poi ripresi del corso della trattazione nei test di ipotesi e nella regressione lineare.



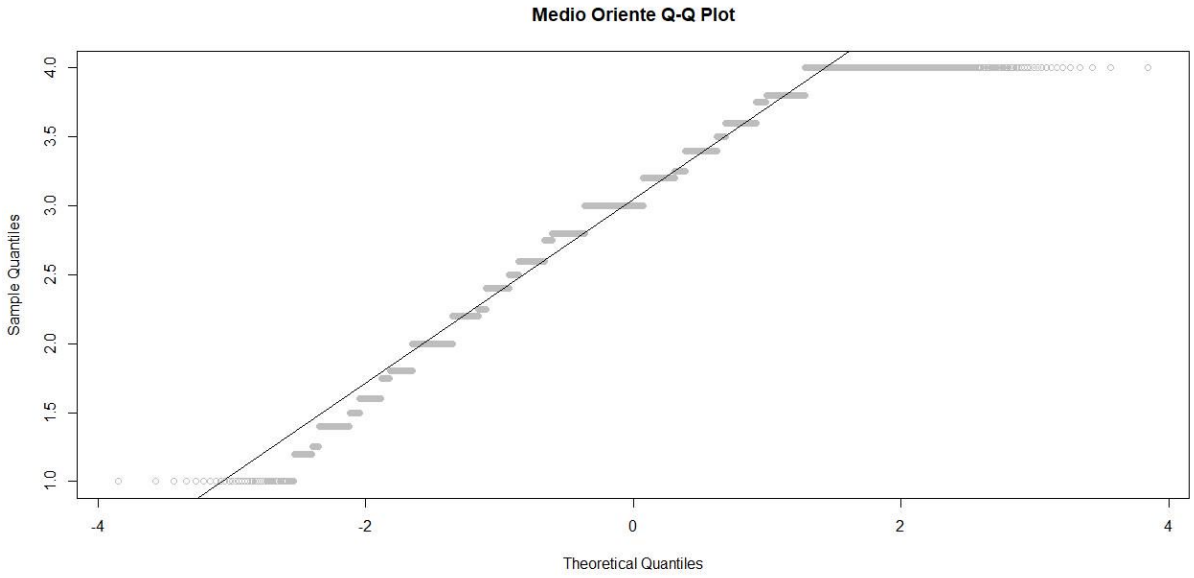
3-Test di normalità

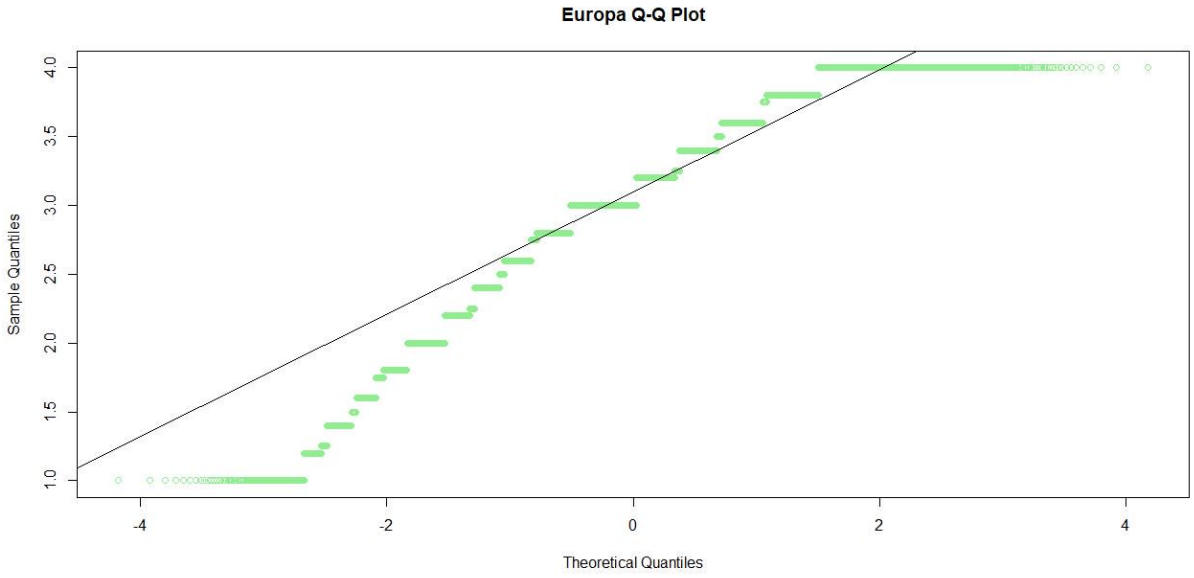
In questa fase ci proponiamo di verificare delle ipotesi precedentemente avanzate sul *Confidence Level*. A tal proposito ci focalizziamo su dei particolari *subset* da noi definiti e li sottoponiamo a dei test di normalità, così da utilizzare in seguito il metodo di verifica delle ipotesi più adatto. In particolare, la normalità è rilevata dai risultati delle seguenti prove:

- *Test di Shapiro-Wilk:* test d'ipotesi che ha come H_0 la normalità del campione.
- *Normal Q-Q Plot:* grafico che confronta la distribuzione cumulata della variabile osservata alla distribuzione della normale.

3.1 Europa e Medio Oriente

Per lo studio della normalità nelle macroaree geografiche, l'elevata numerosità del campione rappresenta un impedimento: il software R non riesce, infatti, ad effettuare uno Shapiro test. Ricorriamo allora al Normal Q-Q plot.

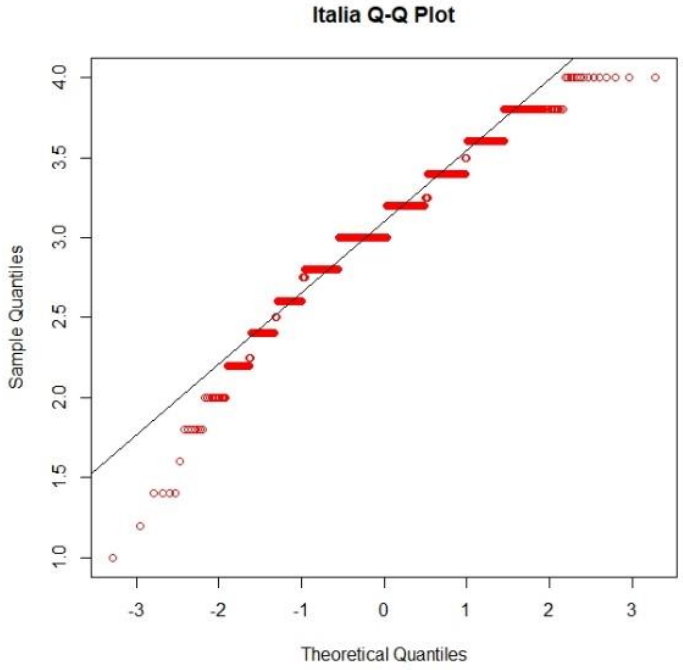




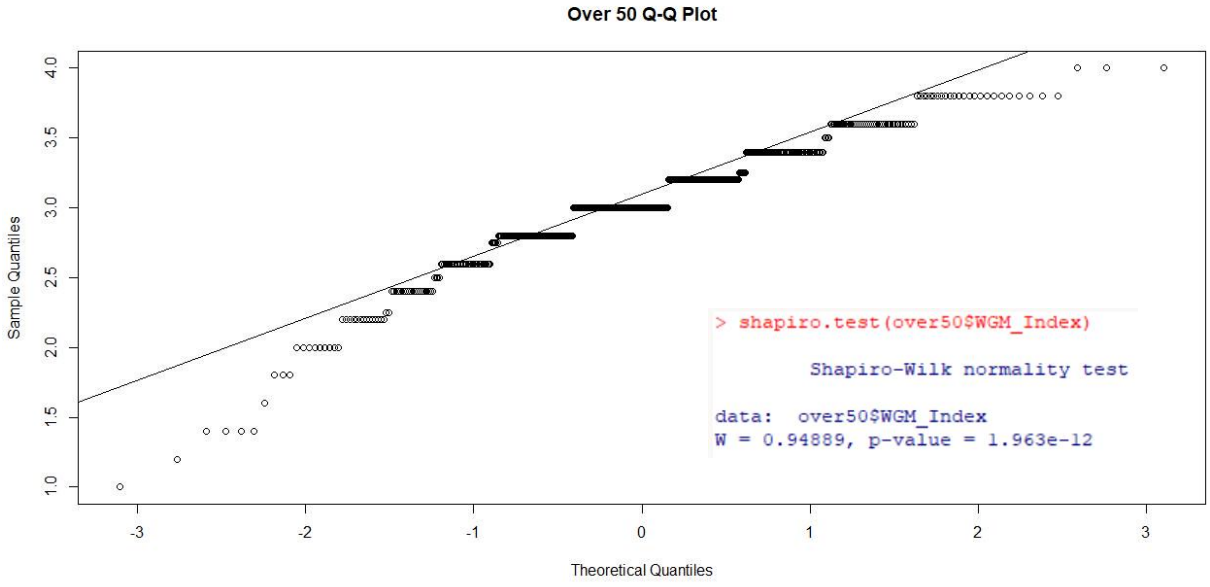
L’aderenza dei dati con la qqline è soltanto parziale (e più accentuata nel Medio Oriente) e va inoltre sfumandosi avvicinandosi alle code, fino a divergere notevolmente. Poiché i dati si discostano in maniera decisa dalla retta, escludiamo la gaussianità di *WGM_Index* nelle due zone. Per di più l’andamento a gradini del grafico è imputabile al fatto che è il nostro indice è un valore discreto.

3.2 Italia

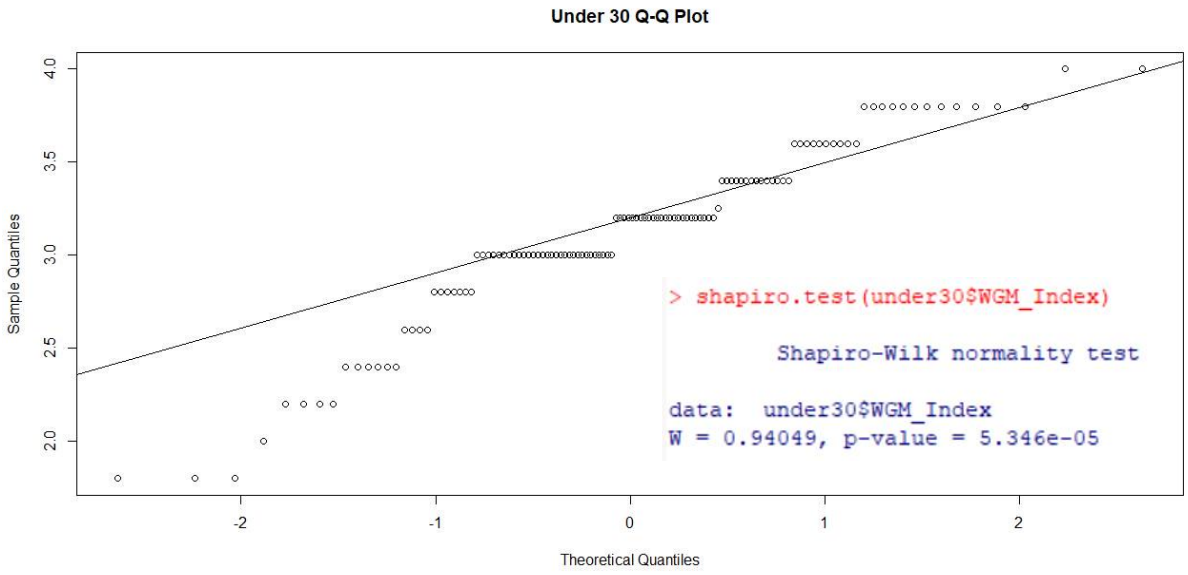
Analogamente cerchiamo di capire se gli stessi indici hanno una distribuzione normale in Italia, sia valutata nel suo complesso sia in sottogruppi definiti, distinti da una caratteristica comune. In particolare, definiamo i subset degli Italiani under30, degli over 50 e di coloro con il grado di istruzione più elevato. Procedendo con le verifiche si riscontrano i seguenti risultati:

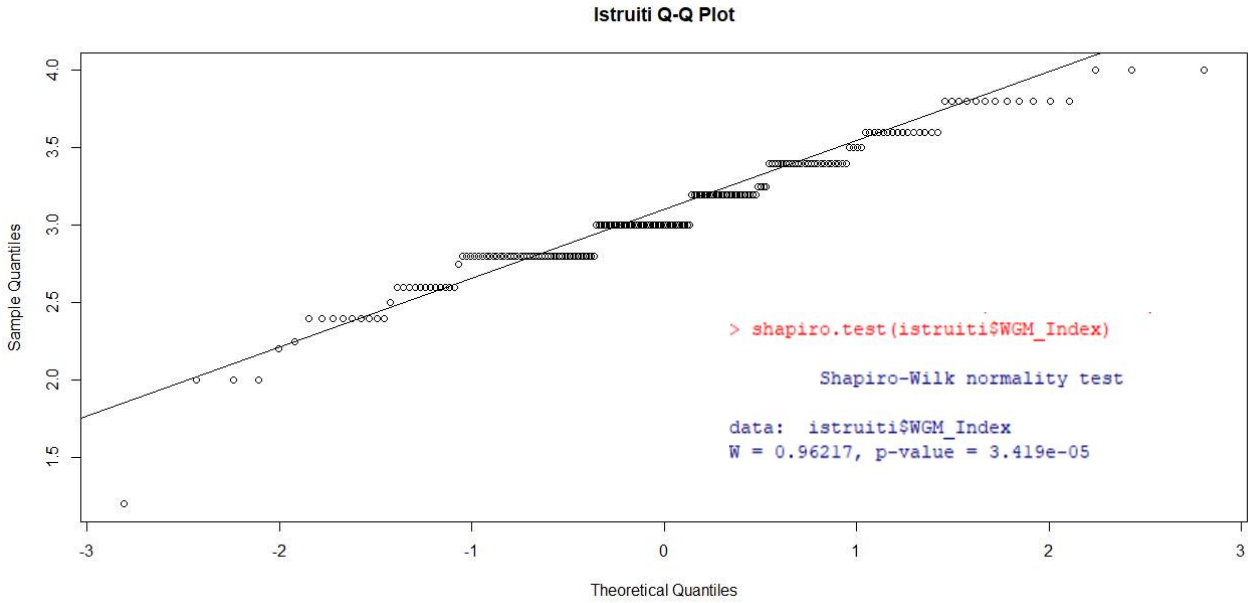



```
        Shapiro-Wilk normality test

data:  it$V1
W = 0.95761, p-value = 5.135e-16
```




```
> shapiro.test(over50$WGM_Index)

        Shapiro-Wilk normality test

data:  over50$WGM_Index
W = 0.94889, p-value = 1.963e-12
```





I dati divergono dalla qqline in maniera ancora più evidente rispetto a quelli degli interi continenti. Questa volta, essendo i campioni inferiori alla soglia limite di numerosità, è possibile eseguire il test di Shapiro-Wilk, dove però otteniamo un *p-value* estremamente basso. Da entrambe le prove si riscontra un'evidenza forte contro l'ipotesi di gaussianità dei dati e siamo pertanto obbligati a rigettarla. Per lo sviluppo delle nostre considerazioni e la realizzazione dei test di ipotesi a seguire, faremo quindi leva sulla numerosità del dataset che presenta centinaia di misure per ogni subset.

4-Test di ipotesi

In questa fase della nostra trattazione, abbiamo cercato di verificare alcune ragionevoli supposizioni sull'andamento dei dati. Partendo da una visione generale del dataset, ci siamo poi focalizzati su zone sempre più circoscritte, con un'attenzione particolare all'Europa e all'Italia. Nell'effettuare i nostri test, abbiamo cercato di ottenere il maggior numero di conclusioni forti possibili, così da rendere minima la probabilità di commettere errori di I specie. Alla luce di ciò, riportiamo i test più significativi, dove si sono poste come ipotesi alternative gli assunti di cui si vuole mostrare l'evidenza. Considerato che nessuna popolazione risulta normale, effettueremo degli Z-test per campioni numerosi, tutti ad un livello di confidenza del 95%.

4.1 Media europea

Innanzitutto vogliamo inserire l'Europa nel contesto globale e capire quanto essa sia coerente con i risultati delle altre regioni. Il primo test effettuato è uno z-test di tipo bilatero e va quindi a verificare se la media europea dell'indice di fiducia si discosta da un valore da noi scelto. Quest'ultimo risulta essere 2.983: la media campionaria globale.
Si avrà quindi:
H_0: $\mu_{Europa} = 2.983$
H_1: $\mu_{Europa} \neq 2.983$

```
> z.test(europa$V2, mu=2.983, sigma.x=sd(europa$V2), conf.level=0.95, alternative="two.sided")

        One-sample z-Test

data:  europa$V2
z = 32.3, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 2.983
95 percent confidence interval:
 3.076664 3.088765
sample estimates:
mean of x
 3.082714
```

Il test presenta un *p-value* infimo, evidenza molto forte della diversità della media Europea da quella globale. Rigettiamo allora H_0 e sviluppiamo questo risultato nel test seguente.

4.2 Differenza tra media europea e medio orientale

Una volta constatata la diversità della media europea, ci occupiamo di confrontarla con un'altra regione del globo per approfondire in che maniera diverga.
Essendo la nostra altra regione sotto esame il Medio Oriente avremo:

H_0: $\mu_{Europa} - \mu_{MedioOriente} \leq 0$
H_1: $\mu_{Europa} - \mu_{MedioOriente} > 0$

```
> z.test(europa$V2, moriente$V2, sigma.x=sd(europa$V2),sigma.y=sd(moriente$V2), conf.level=0.95, alternative="greater")

        Two-sample z-Test

data:  europa$V2 and moriente$V2
z = 5.1109, p-value = 1.603e-07
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 0.02651331         NA
sample estimates:
mean of x mean of y
 3.082714  3.043619
```

Per il valore di *p-value* estratto, è presente una forte evidenza empirica per accettare H_1: l'Europa è pertanto mediamente più fiduciosa nella scienza di quanto non lo sia il Medio Oriente.
Il risultato può essere sottoposto a diverse interpretazioni: il tradizionalismo o la religiosità diffusi nella regione potrebbero rappresentare un freno all'apertura al progresso. D'altro canto, trattandosi di una delle aree più ricche del globo, ci sarà un maggior capitale da poter investire nella ricerca: la disparità nella distribuzione dei beni potrebbe impedire alla maggioranza della popolazione di beneficiare dei risultati delle innovazioni scientifiche.

4.3 Differenza tra medie under30 e over50

Dopo aver analizzato la situazione a livello globale/europeo, prendiamo ora in esame l'Italia. Ci chiediamo pertanto se il grado di fiducia cambi in linea di massima salendo con l'età. Lo sviluppo tecnologico, la diffusione di mezzi informativi rapidi e il più facile accesso all'istruzione degli ultimi decenni, ci fanno supporre che i più giovani siano mediamente più informati e quindi propensi a credere nella scienza. I sottogruppi degli under 30 e over 50 costituiscono rispettivamente il 12.24% e il 53.76% del campione di dati italiani.
Si ha il seguente sistema di ipotesi:
H_0: $\mu_{over50} - \mu_{under30} \geq 0$
H_1: $\mu_{over50} - \mu_{under30} < 0$

```
> z.test(over50$WGM_Index, under30$WGM_Index, sigma.x=sd(over50$WGM_Index), sigma.y=sd(under30$WGM_Index), conf.level=0.95, alternative="less")

        Two-sample z-Test

data:  over50$WGM_Index and under30$WGM_Index
z = -1.987, p-value = 0.02346
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
         NA -0.01632736
sample estimates:
mean of x mean of y
 3.022553  3.117373
```

Il test fornisce un *p-value* pari al 2.346%, dunque la probabilità di commettere errori di I specie risulta bassissima. La conclusione conferma il nostro sospetto iniziale: le nuove generazioni appaiono più propense ad affidarsi alla scienza delle passate.

4.4 Media dei più istruiti

Come ultimo punto ci siamo occupati di indagare come vari l'indice di fiducia del nostro campione in Italia, muovendoci tra il grado di istruzione. In particolare, abbiamo ipotizzato che le persone con un livello d'istruzione più elevato fossero più propense ad affidarsi alla scienza. Infatti, a questa categoria appartengono persone che hanno probabilmente frequentato ambienti di ricerca e sono gli stessi individui che praticano e divulgano la scienza stessa. Per testare il grado di fiducia si è allora fissata una soglia ragionevolmente alta, che ci aspettiamo venga oltrepassata dalla media del campione. La soglia scelta è 2.8 che, rapportata alla nostra scala, corrisponde ad una fiducia del 70%. Precisiamo che in questa fascia si considerano gli individui che hanno conseguito almeno una laurea (oltre i 15 anni di studio), che costituiscono il 20.36% della nostra popolazione.
Impostando il test si ha allora:
H_0: $\mu_{istruiti} \leq 2.8$
H_1: $\mu_{istruiti} > 2.8$

```
> z.test(istruiti$WGM_Index, mu=2.8, sigma.x=sd(istruiti$WGM_Index), conf.level=0.95, alternative="greater")

        One-sample z-Test

data:  istruiti$WGM_Index
z = 8.956, p-value < 2.2e-16
alternative hypothesis: true mean is greater than 2.8
95 percent confidence interval:
 3.020004       NA
sample estimates:
mean of x
   3.0695
```

Dall'elaborazione, il *p-value* del test risulta < $2.2*10^{-16}$. Segue, per un valore così basso, una fortissima evidenza empirica per la validità di H_1. Possiamo decretare con sicurezza che la fiducia media delle persone più istruite in Italia è sopra la soglia prevista.

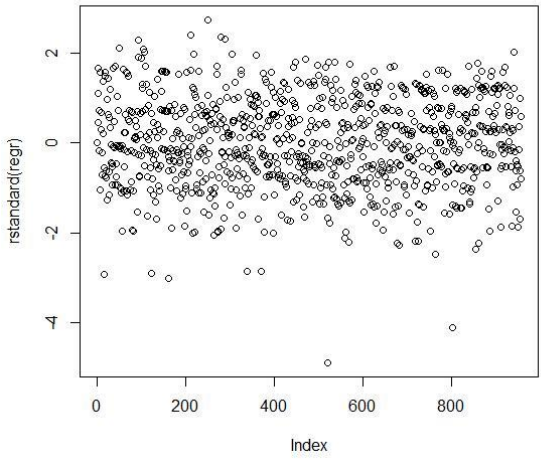
5-Regressione lineare multipla

5.1 Un primo modello

Ci adoperiamo ora per trovare la legge empirica in grado di predire la fiducia nella scienza a partire dagli indici presentati a pagina 4. In particolare, il nostro obiettivo è quello di trovare una correlazione tra la propensione al credere nel progresso e i dati personali riguardanti la vita privata del soggetto.
Cercheremo, quindi, di dedurre un modello di regressione lineare multipla che possa spiegare sufficientemente il *WGM_Index* (*Y*), in funzione di tutti o solo alcuni dei *predittori* presentati (*x1, x2, x3…*). Quasi tutti questi *predittori* possono essere considerati di tipo numerico, assumendo un valore crescente e proporzionale alla *label* che esprimono. In particolare, è una situazione al limite quella di *Employment Status*: decidiamo di interpretarlo come numerico in quanto può essere visto come grado di "inclusione" nel mondo del lavoro e segue quindi anch'esso un ordinamento.
Discorso diverso per *Gender* e *Area type,* che trattiamo come *regressori* categorici: creiamo perciò due *subset* per i valori *uomo/donna* e *urban/rural.* La nostra prima regressione coinvolge tutti gli indici del nostro dataset.

```
Call:
lm(formula = data$WGM_Index ~ ., data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-1.10918 -0.14799 -0.00374  0.16380  0.62337

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        1.932237   0.059083  32.704  < 2e-16 ***
ViewOfScientists   0.688900   0.015975  43.125  < 2e-16 ***
ViewOfScience     -0.091383   0.009008 -10.144  < 2e-16 ***
Age               -0.001698   0.001023  -1.659 0.097364 .
AgeCategories      0.013600   0.023690   0.574 0.566065
Genderuomo         0.005853   0.015021   0.390 0.696890
Education         -0.022990   0.011970  -1.921 0.055069 .
Urban_Ruralurban   0.002959   0.015342   0.193 0.847085
Household_Income   0.003307   0.005660   0.584 0.559146
EMP_2010          -0.012787   0.003828  -3.341 0.000868 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2281 on 947 degrees of freedom
Multiple R-squared:  0.7416,    Adjusted R-squared:  0.7392
F-statistic:   302 on 9 and 947 DF,  p-value: < 2.2e-16
```

Come si può notare dall'output di R, il modello risulta ancora impreciso. In particolare, è evidente la presenza di ben quattro *predittori* poco significativi (*p-value* maggiori del 10%), che andranno esclusi dalla nostra regressione: *AgeCategories, Household_Income* e i due *predittori* categorici.
Per quanto migliorabile, il modello risulta complessivamente buono: i valori di R e R-squared sono soddisfacenti e il modello spiega circa il 74% di *WGM_Index.* Aggiungendo il fatto che il *p-value* dell'F-test è circa 0, possiamo confermare la validità della scelta dei *predittori*.
Tuttavia, è necessario verificare che i residui soddisfino l'ipotesi di gaussianità prima di definire il modello valido. Analizzando lo *scatterplot* dei residui standardizzati, è evidente some essi siano omoschedastici, disposti a nuvola intorno allo 0. La mancanza di *trend* e il fatto che più del 95% dei residui è compreso tra -2 e 2, ci permette di affermarne la normalità. Se ne deduce la validità della nostra regressione.

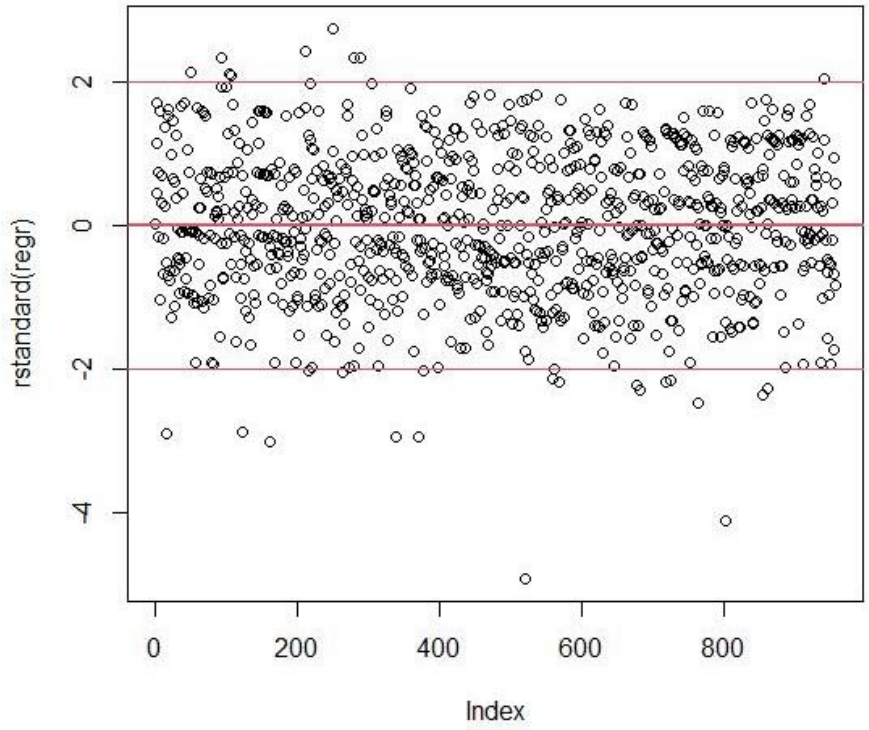


5.2 Modello definitivo

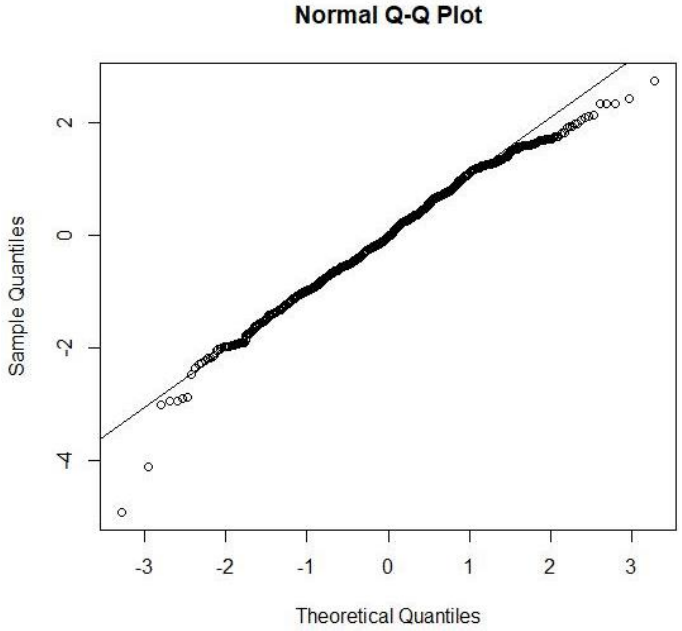
Eseguendo il comando *step* di R (che lavora sulla base dell'*AIC*), effettuiamo successive iterazioni, escludendo uno alla volta i *regressori* non significativi. Il risultato è il miglioramento dello *scatterplot*, mentre non si hanno variazioni importanti negli altri indicatori di bontà.

```
Call:
lm(formula = data$WGM_Index ~ data$ViewOfScientists + data$ViewOfScience +
    data$Age + data$Education + data$EMP_2010)

Residuals:
     Min       1Q   Median       3Q      Max
-1.11312 -0.14904 -0.00564  0.16600  0.62125

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           1.9492115  0.0556687  35.015  < 2e-16 ***
data$ViewOfScientists 0.6895123  0.0159004  43.365  < 2e-16 ***
data$ViewOfScience   -0.0918976  0.0089499 -10.268  < 2e-16 ***
data$Age             -0.0011403  0.0004649  -2.453 0.014343 *
data$Education       -0.0204396  0.0115023  -1.777 0.075887 .
data$EMP_2010        -0.0136052  0.0036996  -3.677 0.000249 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2277 on 951 degrees of freedom
Multiple R-squared:  0.7414,    Adjusted R-squared:  0.7401
F-statistic: 545.3 on 5 and 951 DF,  p-value: < 2.2e-16
```

Analizziamone nuovamente l'output: il valore di *R-squared* è rimasto pressappoco invariato, e risulta sempre più vicino all'*R-squared-adjusted.* Il 74% di quest'ultimo ci tranquillizza sul fatto di non aver commesso *over-fitting*, cioè di aver usato impropriamente i *predittori*. Infatti, l'*R-squared* ha il difetto di aumentare ogni volta che si aggiunge una variabile indipendente al modello (anche se non esplicativa).
Per quanto riguarda la bontà dei *regressori*, basti notare che i *p-value* dei test di significatività singoli di tutti gli indici sono estremamente bassi, tranne il caso di Education che si trova al limite.
Affrontiamo anche questa volta l'analisi dei residui, per stabilirne la gaussianità.
Come evidenzia lo *scatterplot*, i residui standardizzati sono ragionevolmente omoschedastici, e rispettando tutte le condizioni necessarie, così come il precedente. Inoltre, la regressione non sembra variare apprezzabilmente eliminando i due *outliers* di valore vicino al -4. Pertanto, decidiamo di mantenerli per completezza del dataset.



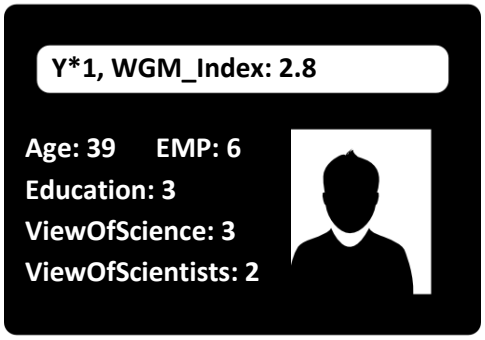
Evidenziamo ora come nel qq-plot la parte centrale aderisca bene alla retta dei quantili: possiamo quindi ragionevolmente considerare i residui gaussiani. In realtà, ciò contraddice il test di Shapiro-Wilk, con un *p-value* eccessivamente basso. E' utile però ricordare che, con un campione troppo numeroso, il test è oltremodo sensibile, e di conseguenza non attendibile. Essendo questo il nostro caso (circa 950 campioni), possiamo pertanto ignorarne il risultato.



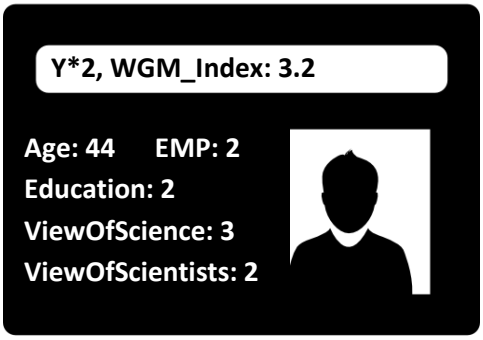

```
        Shapiro-Wilk normality test

data:  rstandard(regr)
W = 0.99001, p-value = 4.283e-06
```

6-Predizioni

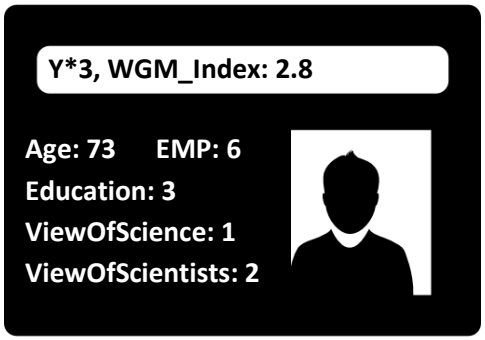
Verifichiamo ora la capacità del modello di predire l'indice di fiducia di misurazioni future Y*, utilizzando dati esclusi dalla regressione e provenienti dallo stesso dataset. Forniamo a R, come input, i valori di *ViewOfScientists*, *ViewOfScience*, *Age*, *Education* e *Employment Status* e confrontiamo i risultati predetti dal nostro modello con quelli effettivi.
Determiniamo, tramite il comando *predict* di R, un intervallo di predizione al livello del 95% per le osservazioni future.



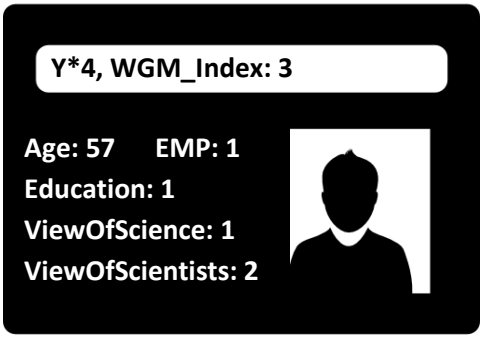
fit	lwr	upr
2.934279	2.486313	3.382245



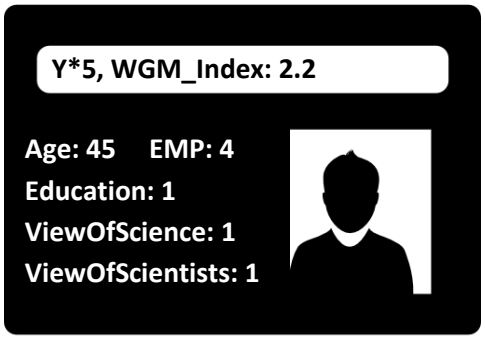
fit	lwr	upr
3.06512	2.615935	3.514306



fit	lwr	upr
3.010144	2.561596	3.458693



fit	lwr	upr
3.137295	2.688681	3.585909



fit	lwr	upr
2.420651	1.97121	2.870092

Per ogni soggetto l'indice reale cade all'interno del nostro intervallo: abbiamo quindi constatato l'attendibilità del modello. Analizziamo per completezza gli errori di predizione tra i valori *fit* predetti e quelli realmente misurati.

```
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-0.2651  -0.2207  -0.2101  -0.1135  -0.1373   0.2657
```

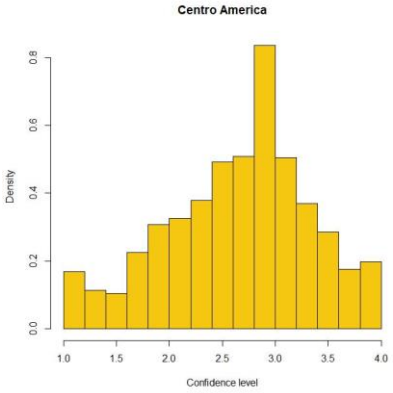
Mediamente il nostro output è leggermente più alto dell'indice effettivo (di 1 decimo), ma possiamo considerarci sufficientemente rassicurati circa la plausibilità delle previsioni.

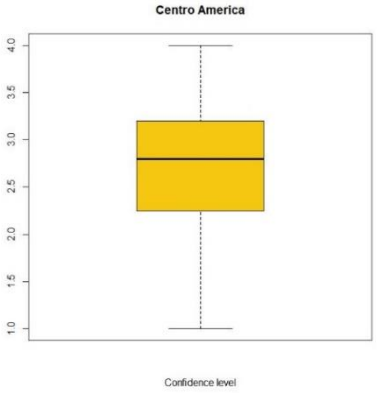
7-Conclusione

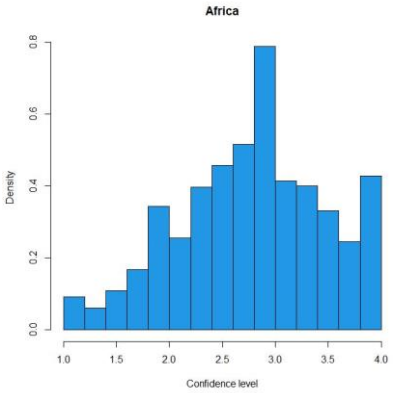
Sebbene non siamo riusciti nello spiegare a pieno come vari l'indice di Fiducia nella scienza (*R-squared* di 0.74 ancora inferiore a 0.8), abbiamo comunque ottenuto risultati rilevanti per le nostre riflessioni.
In particolare, abbiamo dimostrato come la propensione alla fiducia sia soprattutto influenzata dai lavoratori nel settore scientifico, dall'immagine che si percepisce di essi, e dalla loro trasparenza. Un maggiore impegno da parte della comunità scientifica nel divulgare nella maniera più ampia e chiara possibile i propri risultati, porterebbe certamente ad una diminuzione dello scetticismo generale.
Ovviamente, l'altro ruolo fondamentale è rappresentato dalla concezione di scienza. Da come ci viene trasmessa nell'infanzia, al suo impatto nella vita quotidiana, l'indice *ViewOfScience* riassume una visione molto soggettiva, caratterizzata da numerose sfumature.
E' inoltre interessante la significatività dell'*Employment Status:* una persona che vive isolata dalla società può falsamente incolpare della sua improduttività il progresso tecnologico.
Analogamente per l'*età,* spesso gli anziani tendono ad essere più tradizionalisti e meno propensi ad accogliere le novità o sono generalmente meno attratti dalle notizie di attualità.
Diversamente, un *predittore* che ci aspettavamo essere di più largo impatto è l'*educazione.* Tendenzialmente si reputano le persone più istruite come più aperte al futuro e più in grado di dare giudizi ponderati e oggettivi sulle informazioni ricevute dal mondo esterno. Evidentemente le cose non vanno sempre così, o perlomeno la strada da seguire è leggermente diversa: probabilmente, con un indice diviso anche per tipologia di studio, si sarebbero ottenuti risultati più rilevanti.
Il concetto che abbiamo analizzato è complesso e difficile da concretizzare in valori numerici. Spesso la fiducia è legata ad esperienze personali vissute, a traumi o momenti positivi che possono offuscare il nostro giudizio. La percentuale di variabilità di *Y* che non siamo riusciti a spiegare dipende perciò da dati non inclusi nel dataset. Si sarebbero ottenuti risultati più accurati se si fossero presi in considerazione nel questionario altri aspetti della vita del singolo, come ad esempio l'uso dei social, l'esposizione alle fake news o addirittura il quoziente intellettivo.

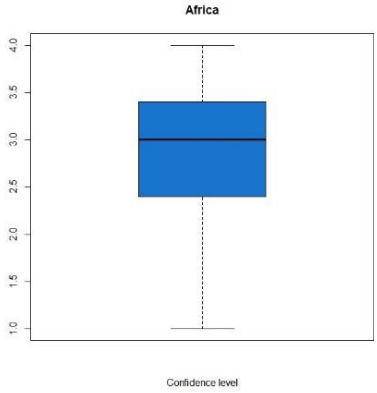
8- Grafici

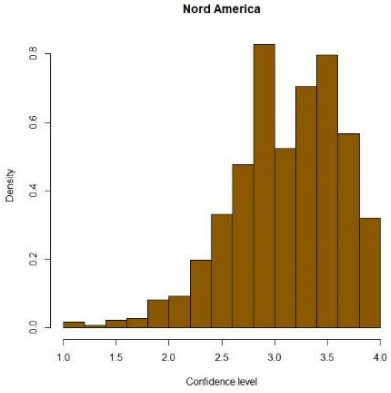
8.1 Confidence level

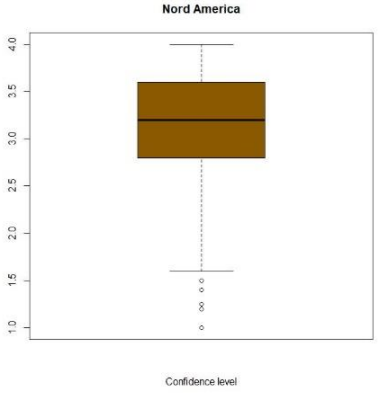


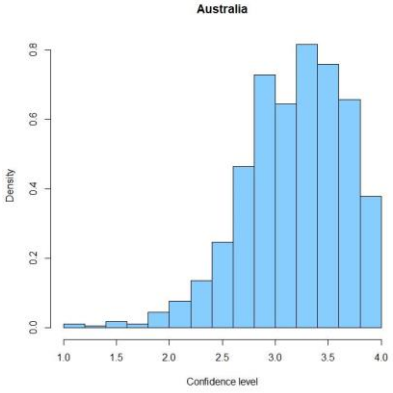


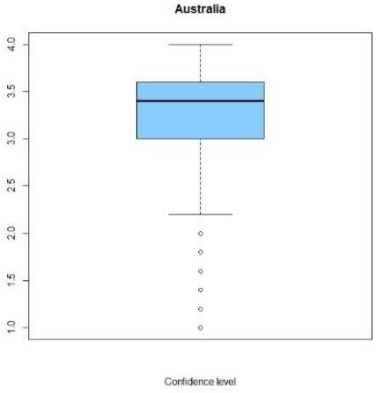


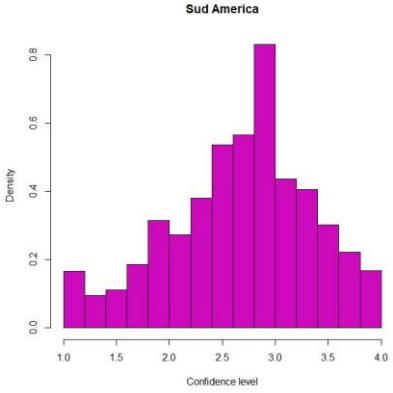


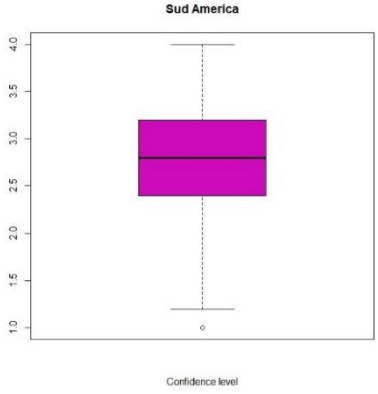


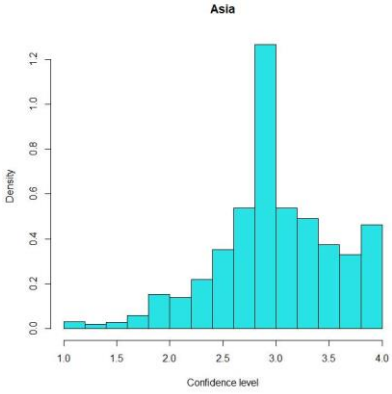


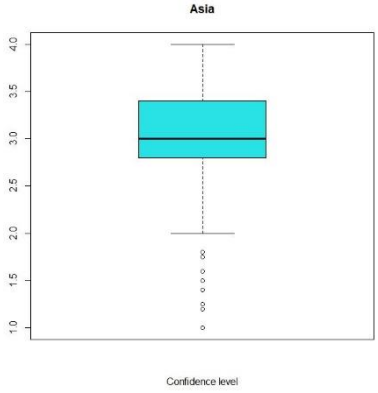






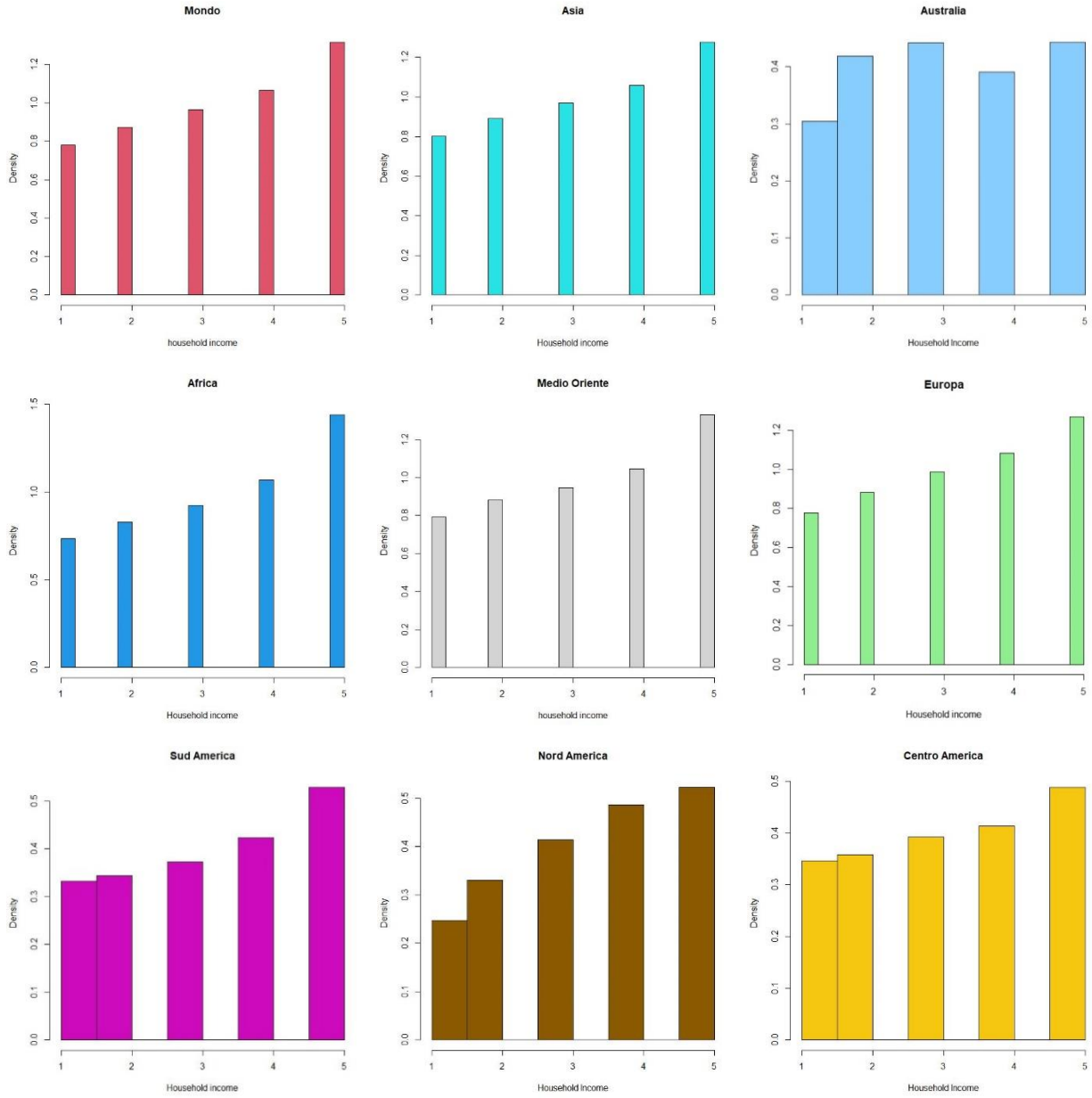




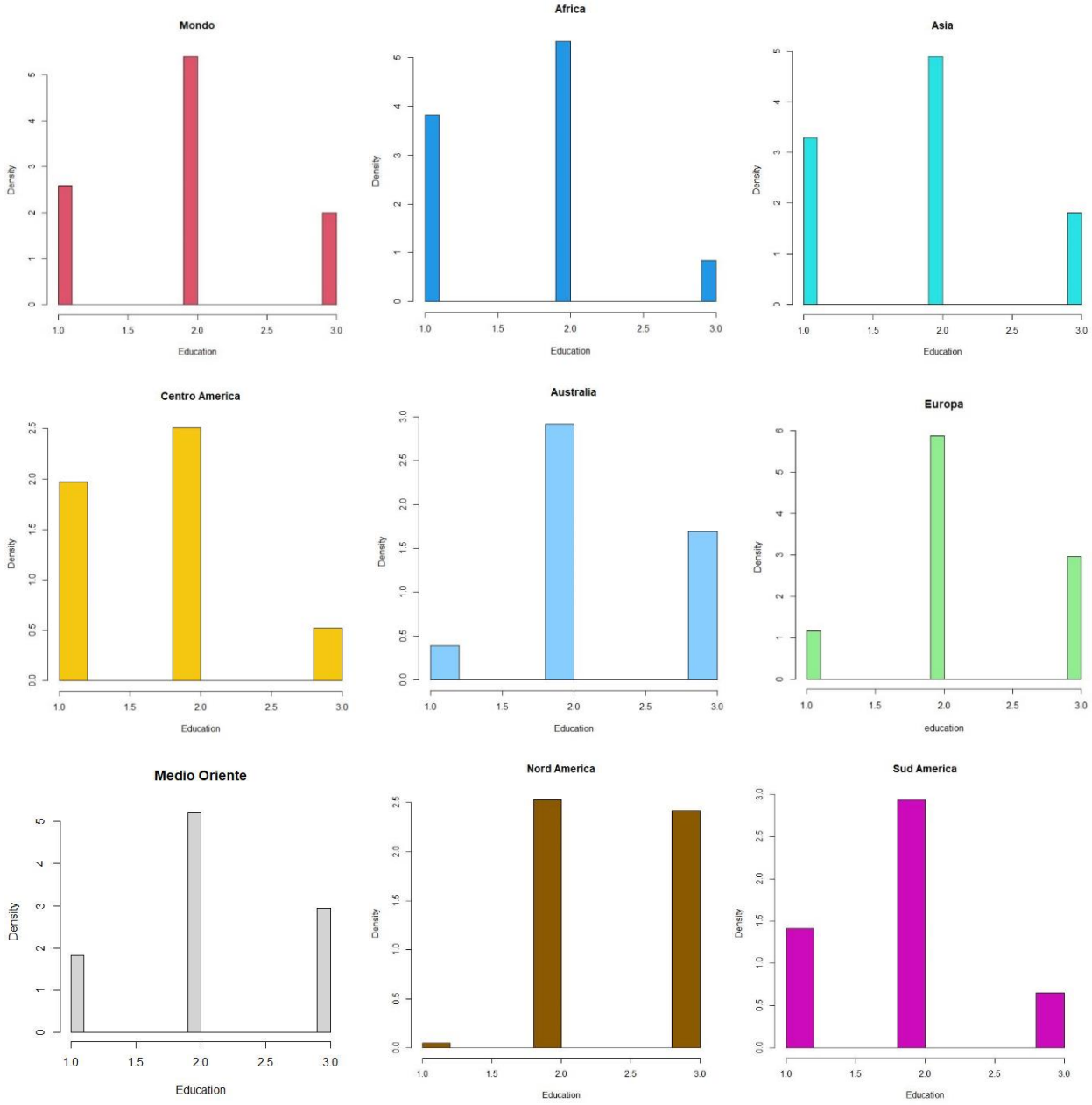


8.2 Household Income

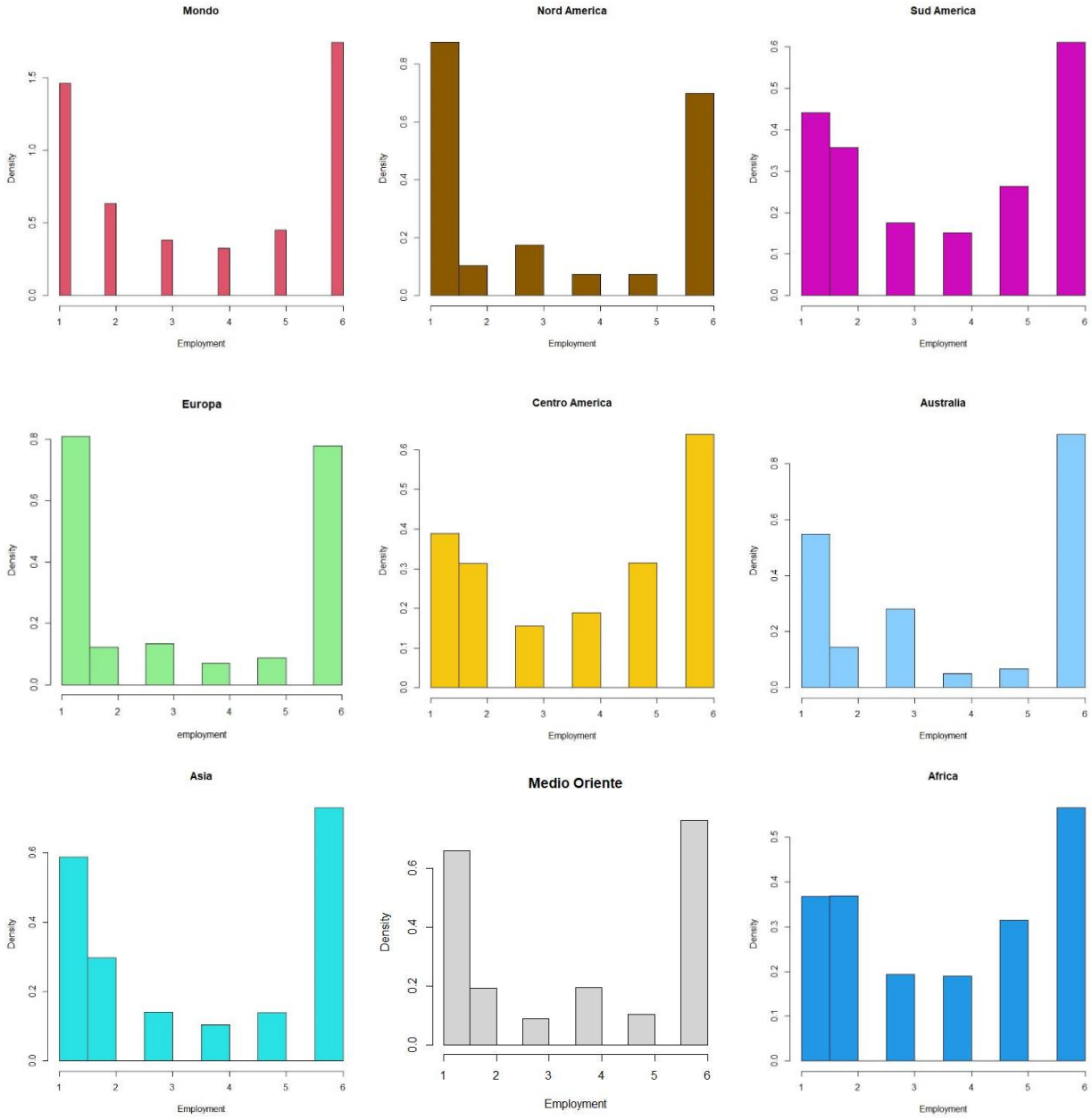
N.B. per i successivi indici, essendo discreti, utilizzeremo dei grafici a barre. Le colonne sono centrate sui valori interi, e rappresentano la propria label.



8.3 Education



8.4 Employment status



9- Voci Correlate

- Dataframe: https://beta.ukdataservice.ac.uk/datacatalogue/studies/study?id=8466#!/details
- (1*) https://www.tuttitalia.it/statistiche/popolazione-eta-sesso-stato-civile-2019/
- Supporto computazionale:
  https://r-statistics.co/ggplot2-cheatsheet.html
  https://www.statmethods.net/index.html
- Bibliografia: Montgomery Runger Hubele, *Engineering Statistics*, 5th edition, 2010