Analisi di alcune distribuzioni nell'ambito delle professioni giudiziarie

Giorgio Spugnesi – Metodi della Fisica per le Scienze Umane – a.a. 2012/2013

Oggetto dell'analisi e considerazioni generali

Nello studiare le distribuzioni in frequenza di dati relativi a campioni più o meno vasti della popolazione umana si è soliti attendersi due possibili tipologie di distribuzione: quella dotata di scala e quella priva di scala. Alcune distribuzioni, come l'altezza o l'età al momento del matrimonio, rispondono infatti a distribuzioni che seguono curve di tipo demografico (dette anche “a campana”) o, più precisamente, rispondenti alle curve di Gompertz, mentre altre, come quella dei cognomi, a distribuzioni che seguono la cosiddetta power law.
Il presente lavoro si pone lo scopo di analizzare alcune distribuzioni all'interno di due particolari campioni di popolazione: gli avvocati iscritti all'Ordine di Roma e il personale (amministrativo e giudiziario) in servizio presso alcuni degli Uffici Giudiziari dello Stato Italiano.
Per entrambi i campioni sarà studiata sia la distribuzione in frequenza dei cognomi sia l'età media di inizio della professione (considerando l'iscrizione all'Albo per gli avvocati e la nomina ministeriale per il personale giudiziario).
Lo studio della distribuzione in frequenza dei cognomi all'interno di un campione della popolazione gode ormai di una ricca letteratura e, lungi dall'essere un esercizio fine a se stesso, fornisce preziose informazioni a numerose discipline sia umanistiche (storia, analisi demografica e sociologica) che scientifiche (genetica, studio dell'evoluzione). Proprio in questo contesto, anzi, i confini tra scientifico e umanistico vengono a scomparire ed emerge un approccio prevalente che, partendo dalla lettura di dati quantitativi, è in grado di fornire contributi qualitativi alle varie discipline.
L'analisi dell'età di inizio attività, condotta distinguendo sia per sesso, sia per decennio di nascita, fornisce dati interessanti ad un'analisi di tipo sociologico sulla facilità di accesso alla professione anche in un'ottica di evoluzione temporale. In questo contesto, quindi, ancora una volta, l'analisi di dati quantitativi permette una interpretazione qualitativa e trasversale tra discipline.

Questioni metodologiche

Sebbene le due analisi siano state condotte separatamente, distingueremo per chiarezza i due campioni, nominando campione A quello relativo agli avvocati e campione B quello relativo al personale degli uffici giudiziari.
Entrambi i campioni provengono da basi dati relazionali facenti parte di sistemi informatici gestionali. Non è stato effettuato alcun tipo di controllo sui dati pertanto eventuali anomalie saranno discusse in fase di analisi.
La base dati del campione A conta 22777 records mentre quella del campione B 14798 records.
Per quanto riguarda il campione B, solamente 3580 records forniscono dati in merito alle date di nascita e di inizio attività pertanto le analisi sul reclutamento sono state condotte su un campione ridotto.
I dati relativi ai due campioni sono stati estratti dalle rispettive basi dati ed importati in un database MySQL in modo da poter essere interrogati attraverso il linguaggio SQL ed ottenere dati quantitativi e di frequenza. Le informazioni così estratte sono state elaborate attraverso il software di calcolo numerico Octave che ha permesso anche la creazione di grafici.
Sulla base dei dati numerici e dei grafici, infine, è stata condotta l'analisi e sono state tratte le conclusioni che sono oggetto di questo lavoro.
Ritengo utile, quindi, data anche la natura didattica del lavoro, soffermarmi sui vari procedimenti di estrazione ed elaborazione dei dati.

Estrazione dei dati

Per quanto riguarda la distribuzione in frequenza dei cognomi, sia per il campione A che per il campione B, si è proceduto nel modo seguente.
Dato che la base dati contiene una riga (record) per ciascun soggetto e che uno dei campi del record è il cognome, in prima istanza è possibile determinare le occorrenze di ciascun cognome:

```
SELECT COUNT(cognome) AS o FROM tabella GROUP BY cognome
```

Notare il raggruppamento sul cognome che permette di sommare le occorrenze di cognomi uguali.
Useremo adesso questa prima selezione come base su cui contare la frequenza delle occorrenze, ovvero quanti cognomi sono presenti una, due, tre, ecc. volte.

```
SELECT o, COUNT(o) AS f FROM (SELECT COUNT(cognome) AS o FROM [tabella]
GROUP BY cognome) AS tbl GROUP BY o ORDER BY o DESC
```

In questo modo si selezionano le occorrenze (raggruppandole) e le relative frequenze ottenendo una tabella ordinata per numero decrescente di occorrenze.

Poiché le due tabelle relative ai campioni sono, almeno per questa parte, simili (hanno entrambe un campo “cognome”), lanciando l'istruzione SQL mostrata sopra, con l'opportuno nome al posto di [tabella], si ottengono i seguenti risultati.

Campione A

o	f
87	1
67	1
51	2
46	1
42	1
40	1
36	1
35	1
33	2
32	1
29	1
28	1
27	3
25	3
24	3
23	3
22	6
21	4
20	4
19	7
18	5
17	6
16	6
15	8
14	11
13	25
12	20
11	33
10	31
9	40
8	71
7	91
6	160
5	248
4	418
3	794
2	2026
1	7770

Campione B

o	f
71	1
60	1
47	1
29	2
27	2
25	1
22	2
20	3
19	1
18	3
17	2
16	4
15	10
14	5
13	3
12	6
11	17
10	19
9	31
8	39
7	49
6	70
5	131
4	205
3	389
2	1281
1	6942

Da notare che, mentre nella prima istruzione SQL sarebbe stato possibile mostrare anche l'elenco dei cognomi oltre all'occorrenza (e sapere in questo modo quale sia il cognome più presente nella base dati), nella seconda istruzione, dovendo compiere un'ulteriore raggruppamento di dati per contare la frequenza delle occorrenze, viene perduto qualsiasi dato relativo al singolo oggetto di indagine.

Più complesso si presenta il calcolo dell'età di inizio professione, soprattutto per il campione A nella cui tabella le date sono conservate non in formato data ma come semplici stringhe. MySQL quindi non è in grado di effettuare su di esse operazioni di differenza a meno di non ricorrere a funzioni di conversione. Per calcolare l'età di inizio professione si procede facendo la sottrazione tra l'anno di inizio professione e l'anno di nascita e sottraendo 1 qualora giorno e mese di nascita siano successive, rispetto all'anno solare, a giorno e mese di inizio professione. Questa formula può essere scritta

anno(data inizio) - anno(data nascita) - (giorno e mese inizio < giorno e mese nascita)

in considerazione del fatto che il confronto produrrà un esito booleano che sarà interpretato da MySQL come 0 (falso) o 1 (vero) andando a togliere tale valore dal risultato della sottrazione iniziale.
Con tutte le conversioni necessarie (alcune esclusivamente per il campione A), l'istruzione SQL per ricavare l'età di inizio attività per ciascun soggetto del campione è la seguente:

```
SELECT YEAR(DATE_FORMAT(STR_TO_DATE([dataInizio], '%d.%m.%Y'), '%Y-%m-
%d')) - YEAR(DATE_FORMAT(STR_TO_DATE([dataNascita], '%d.%m.%Y'), '%Y-%m-
%d')) - (DATE_FORMAT(STR_TO_DATE([dataInizio], '%d.%m.%Y'), '%m%d') <
DATE_FORMAT(STR_TO_DATE([dataNascita], '%d.%m.%Y'), '%m%d')) AS eta
FROM [tabella]
```

Tali risultati costituiscono la base su cui andare a calcolare le varie distribuzioni indicate in apertura del lavoro. A tale scopo, risulta opportuno aggiungere ai singoli record i valori ottenuti come nuovo campo. In questo modo si semplificano le successive operazioni di estrazione dei dati.
Seguendo un procedimento analogo a quanto fatto per la distribuzione dei cognomi, si può calcolare l'occorrenza di ciascuna età, filtrando eventualmente su altri campi presenti nella tabella come il sesso. Questa informazione è presente nella tabella del campione A come due record separati (maschio con valore stringa = X, femmina con valore stringa = X), soluzione certamente non ottimale, mentre per il campione B è stato adoperato un solo campo con due possibili valori stringa (M, F).
Infine, volendo analizzare l'età di inizio attività per decennio di nascita, il filtro è stato ripetuto selezionando solo i records relativi a soggetti nati dopo una certa data e prima di un'altra, individuando quindi una serie di sottocampioni i cui dati sono stati confrontati tra loro.
Per poter disporre di dati importabili nel software Octave, si è scelto di estrarre le serie di dati su file di testo, uno per ciascuna colonna di risultati. Il processo di esecuzione delle istruzioni SQL è stato quindi automatizzato in modo da riversare l'output su file. Per quanto riguarda i nomi dei file, è stato stabilito il seguente standard:

X-*n*-*y* .txt

X : nome del campione <A,B>
n : numero assegnato all'interrogazione
y : colonna dati <a,b>

e ad ogni interrogazione alle basi dati è stato assegnato un numero:

1) frequenza cognomi;
2) frequenza età;
3) frequenza età per sesso (maschi);
4) frequenza età per sesso (femmine);
5) frequenza età per anno nascita (1940-1950) ;
6) frequenza età per anno nascita (1950-1960) ;
7) frequenza età per anno nascita (1960-1970) ;
8) frequenza età per anno nascita (1970-1980) ;
9) frequenza età per anno nascita (1980-1990) .

In questo modo, attraverso il nome del file, è possibile determinare il tipo dei dati.

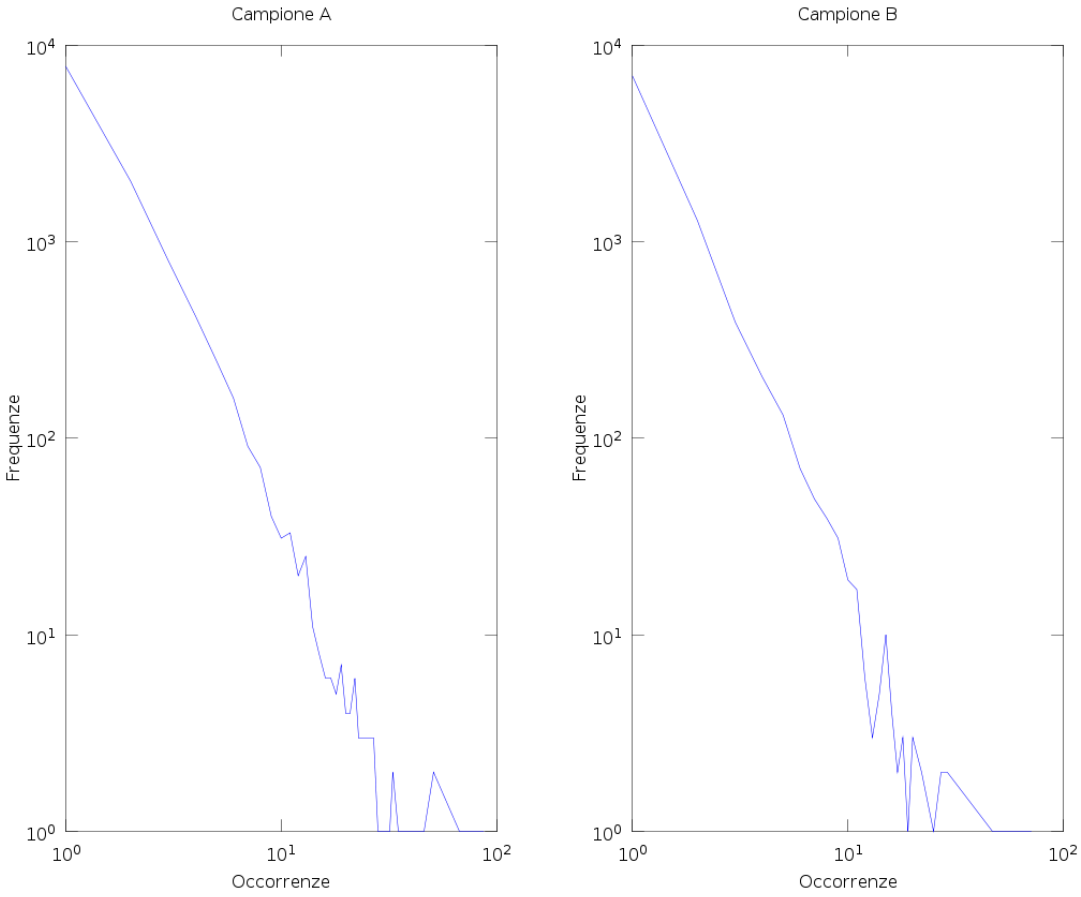
Elaborazione dei dati e generazione dei grafici

Come già ricordato, Octave è una applicazione software per l'analisi numerica e, con l'aggiunta del componente GNUPlot, per la generazione di grafici. Il tipo di dati prevalentemente usato è la matrice (e di conseguenza il vettore come matrice monodimensionale); è pertanto possibile definire delle variabili all'interno dell'ambiente contenenti ennuple (vettori appunto) composte dai risultati delle interrogazioni alla base dati.
Le serie di dati contenuti nei vari file sono state assegnate ad altrettante variabili, chiamate come il nome del file stesso (senza trattino separatore) e definite all'interno di uno script in modo da poter creare l'ambiente di analisi con un semplice comando.
Per ogni set di dati, dopo aver determinato la rappresentazione grafica più idonea, è stato creato uno script che genera il grafico definendo la scala opportuna, le etichette e tutte le altre informazioni necessarie. I grafici generati sono stati infine inseriti nel presente lavoro.

Analisi dei dati

Distribuzione dei cognomi

Osservando le serie numeriche dell'analisi 1 (distribuzione dei cognomi) appare evidente una distribuzione logaritmica della frequenza. Gli studi sui cognomi hanno sempre identificato una distribuzione rispondente alla power law ed anche in questo caso possiamo aspettarci una tale distribuzione. Osservando il grafico generato sia per il campione A che per il campione B, notiamo che la distribuzione segue la power law anche se con un notevole rumore nella coda.



Definendo la formula[1] della power law come

$$p(x)=Cx^{-a}$$

possiamo calcolare, per entrambi i campioni, l'esponente

$$a=1+n[\sum_{1}^{n}\ln(x_i/x_{min})]^{-1}$$

e l'errore statistico

$$s=(a-1)/\sqrt{n}$$

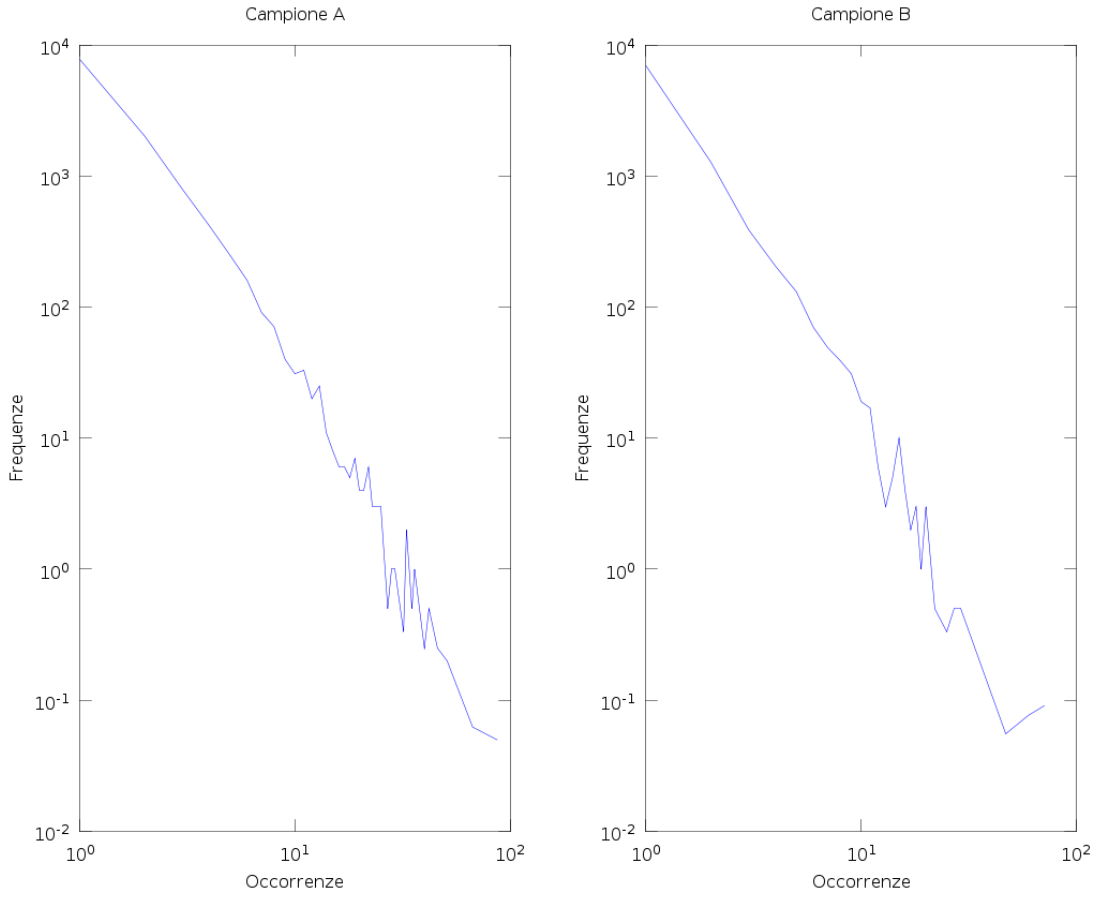
I valori calcolati per il campione A risultano 1.4346 ± 0.0705 e per il campione B 1.4070 ± 0.0783.

Come si può osservare, la presenza di interruzioni nella serie delle occorrenze genera un notevole rumore nella coda e l'appiattimento del grafico sull'ascissa. Trattandosi di misura di frequenze, è possibile ridefinire il grafico calcolando, per i valori successivi ad ogni interruzione, la probabilità del valore misurato con la formula

1 Per gli aspetti matematici della power law si fa riferimento a M.E.J. Newmann *Power laws. Pareto distributions and Zipf's law.* Contemporary Physics 46, 323 – 351 (2005)

$$p(x)=1/(x_1-x_0)$$

dove x_1 è il valore dell'occorrenza successivo all'interruzione e x_0 quello precedente.
Il grafico tracciato a partire dai valori corretti, pur non essendo privo di interruzioni della linearità, mostra comunque il proseguimento della distribuzione lungo la retta anche per i valori con occorrenza più alta.



Per l'analisi dell'isonimia si è fatto ricorso all'elenco dei primi 100 cognomi più presenti in Italia, reperito tramite una ricerca in rete[2]. Purtroppo l'attendibilità e la provenienza originaria dei dati non sono verificabili.
Questo elenco è stato confrontato con quello dei primi 100 cognomi per occorrenza di ciascun campione. Per poter assegnare un valore numerico a ciascun cognome si è confrontato la posizione in classifica di ciascun cognome del campione con quella nella lista nazionale, calcolando la differenza tra le due posizioni. Ai cognomi non presenti nella lista nazionale si è assegnato il valore 100.

Tabella 1 – I primi cento cognomi confrontati con quelli dei due campioni

Pos.	**Elenco nazionale**	**Campione A**	**Campione B**
1	ROSSI	ROSSI	RUSSO
2	RUSSO	RUSSO	ESPOSITO
3	FERRARI	MANCINI	ROSSI
4	ESPOSITO	ROMANO	DE LUCA
5	BIANCHI	RICCI	GIORDANO

2 Fonte: http://www.cognomix.it/top100_cognomi_italia.php (consultato il 18/03/2013)

6	ROMANO	DE ANGELIS	MARINO
7	COLOMBO	GRECO	ROMANO
8	RICCI	GIORDANO	GALLO
9	MARINO	LOMBARDI	BRUNO
10	GRECO	GENTILE	LOMBARDI
11	BRUNO	DE LUCA	RUGGIERO
12	GALLO	CONTI	SANTORO
13	CONTI	DE SANTIS	FERRARA
14	DE LUCA	BIANCHI	COPPOLA
15	MANCINI	MARINO	PEPE
16	COSTA	MARINI	DE ROSA
17	GIORDANO	MONACO	PALUMBO
18	RIZZO	ESPOSITO	GRECO
19	LOMBARDI	LONGO	CONTE
20	MORETTI	CARUSO	RICCI
21	BARBIERI	BRUNO	RIZZO
22	FONTANA	D'AMICO	DE ANGELIS
23	SANTORO	SILVESTRI	PELLEGRINO
24	MARIANI	PROIETTI	VITALE
25	RINALDI	LEONE	BIANCHI
26	CARUSO	FERRI	MORETTI
27	FERRARA	PETRUCCI	SORRENTINO
28	GALLI	VITALE	AMATO
29	MARTINI	GALLO	SIMONETTI
30	LEONE	RINALDI	FERRARO
31	LONGO	D'ANGELO	DE MARTINO
32	GENTILE	PELLEGRINI	GENTILE
33	MARTINELLI	ROMEO	CAPUTO
34	VITALE	GIULIANI	FERRARI
35	LOMBARDO	SANTORO	SESSA
36	SERRA	FERRARI	FIORE
37	COPPOLA	NERI	LIGUORI
38	DE SANTIS	MARIANI	D'AURIA
39	D'ANGELO	BERNARDINI	COZZOLINO
40	MARCHETTI	MATTEI	MELE
41	PARISI	DE SIMONE	D'ANGELO
42	VILLA	MORELLI	MANCINI
43	CONTE	FUSCO	BORRELLI
44	FERRARO	CARBONE	CUOMO
45	FERRI	CONTE	SCOGNAMIGLIO
46	FABBRI	NAPOLITANO	GUARINO
47	BIANCO	TESTA	D'AMICO
48	MARINI	RIZZO	MONACO
49	GRASSO	D'AGOSTINO	LEONE
50	VALENTINI	PELLEGRINO	BALDI
51	MESSINA	PUCCI	CATTANEO
52	SALA	PAOLETTI	GRASSO
53	DE ANGELIS	COSTANTINI	DE SIMONE
54	GATTI	GRIMALDI	GUIDA
55	PELLEGRINI	MORETTI	TUCCI
56	PALUMBO	IZZO	BASILE
57	SANNA	MAZZA	PALMIERI
58	FARINA	SERRA	D'AMBROSIO
59	RIZZI	PARISI	MORELLI
60	MONTI	ANGELINI	AMOROSO
61	CATTANEO	DE STEFANO	CARBONE
62	MORELLI	ORLANDO	GRIMALDI
63	AMATO	GUERRA	DE STEFANO

64	SILVESTRI	NATALE	NAPOLITANO
65	MAZZA	AMATO	BEVILACQUA
66	TESTA	GENTILI	FALCONE
67	GRASSI	PACE	BARBATO
68	PELLEGRINO	FIORE	FERRO
69	CARBONE	BARONE	DI MARTINO
70	GIULIANI	DE MARCO	CALABRESE
71	BENEDETTI	DE ROSA	IZZO
72	BARONE	PUGLIESE	NAPOLI
73	ROSSETTI	BUCCI	FONTANA
74	CAPUTO	BENEDETTI	PAGANO
75	MONTANARI	ANTONELLI	GRIECO
76	GUERRA	RUGGIERO	NERI
77	PALMIERI	VALENTINI	RICCIARDI
78	BERNARDI	BLASI	MARINI
79	MARTINO	SALERNO	GALLI
80	FIORE	BONANNI	CONTI
81	DE ROSA	GIANNINI	D'ELIA
82	FERRETTI	LOMBARDO	D'AGOSTINO
83	BELLINI	PALOMBI	MANTOVANI
84	BASILE	GRILLO	MARCHETTI
85	RIVA	GROSSI	SCOTTI
86	DONATI	RICCIARDI	PARISI
87	PIRAS	SANTINI	DE SANTIS
88	VITALI	TUCCI	COLOMBO
89	BATTAGLIA	VALENTI	MARTINELLI
90	SARTORI	ANTONINI	MARIANI
91	NERI	CARDARELLI	ALBANESE
92	COSTANTINI	PINTO	FIORILLO
93	MILANI	TURCO	MAZZA
94	PAGANO	VITALI	IORIO
95	RUGGIERO	ALBANESE	GRAZIANO
96	SORRENTINO	FERRARO	DI MAIO
97	D'AMICO	FRANCO	BARBIERI
98	ORLANDO	GRASSI	MESSINA
99	DAMICO	OLIVA	RINALDI
100	NEGRI	PALERMO	BENEDETTI

Tabella 2 – La posizione relativa al campione nazionale per i cognomi dei due campioni

Campione A		**Campione B**	
FERRARO	-52	COLOMBO	-81
LOMBARDO	-47	BARBIERI	-76
MORETTI	-35	RINALDI	-74
FERRARI	-33	CONTI	-67
GRASSI	-31	MARIANI	-66
RIZZO	-30	MARTINELLI	-56
VALENTINI	-27	FONTANA	-51
SERRA	-22	GALLI	-51
PARISI	-18	DE SANTIS	-49
GALLO	-17	MESSINA	-47
MARIANI	-14	PARISI	-45
ESPOSITO	-14	MARCHETTI	-44
SANTORO	-12	FERRARI	-31
BRUNO	-10	MARINI	-30
BIANCHI	-9	BENEDETTI	-29
VITALI	-6	MAZZA	-28

MARINO	-6	MANCINI	-27
RINALDI	-5	BIANCHI	-20
BENEDETTI	-3	LEONE	-19
CONTE	-2	RICCI	-12
AMATO	-2	GRECO	-8
ROSSI	0	MORETTI	-6
RUSSO	0	GRASSO	-3
CONTI	1	RIZZO	-3
ROMANO	2	ROSSI	-2
GRECO	3	DANGELO	-2
DE LUCA	3	ROMANO	-1
BARONE	3	GENTILE	0
RICCI	3	RUSSO	1
LEONE	5	ESPOSITO	2
VITALE	6	BRUNO	2
CARUSO	6	MORELLI	3
DANGELO	8	MARINO	3
MAZZA	8	GALLO	4
GIORDANO	9	CARBONE	8
DE ROSA	10	LOMBARDI	9
LOMBARDI	10	CATTANEO	10
FIORE	12	VITALE	10
LONGO	12	DE LUCA	10
MANCINI	12	SANTORO	11
GUERRA	13	GIORDANO	12
PELLEGRINO	18	FERRARO	14
TESTA	19	FERRARA	14
FERRI	19	NERI	15
RUGGIERO	19	PAGANO	20
MORELLI	20	PALMIERI	20
GENTILE	22	COPPOLA	23
PELLEGRINI	23	CONTE	24
CARBONE	25	BASILE	28
DE SANTIS	25	DE ANGELIS	31
MARINI	32	AMATO	35
ORLANDO	36	PALUMBO	39
GIULIANI	36	CAPUTO	41
COSTANTINI	39	FIORE	44
SILVESTRI	41	PELLEGRINO	45
DE ANGELIS	47	DAMICO	52
NERI	54	DE ROSA	65
DAMICO	77	SORRENTINO	69
GRILLO	100	RUGGIERO	84
SALERNO	100	BEVILACQUA	100
PAOLETTI	100	DELIA	100
GROSSI	100	RICCIARDI	100
BUCCI	100	DI MARTINO	100
RICCIARDI	100	DAGOSTINO	100
DE SIMONE	100	SCOGNAMIGLIO	100
FRANCO	100	BARBATO	100
PACE	100	CUOMO	100
DAGOSTINO	100	FALCONE	100
FUSCO	100	DAMBROSIO	100
GIANNINI	100	ALBANESE	100
PUCCI	100	PEPE	100
ALBANESE	100	FERRO	100
PROIETTI	100	GUIDA	100
OLIVA	100	SCOTTI	100

MONACO	100	GRAZIANO	100
PUGLIESE	100	GUARINO	100
PINTO	100	MONACO	100
ANGELINI	100	MANTOVANI	100
TUCCI	100	DAURIA	100
PALOMBI	100	TUCCI	100
ANTONINI	100	DE STEFANO	100
CARDARELLI	100	SIMONETTI	100
DE STEFANO	100	COZZOLINO	100
ROMEO	100	GRIMALDI	100
SANTINI	100	DE SIMONE	100
PALERMO	100	FIORILLO	100
TURCO	100	IZZO	100
NAPOLITANO	100	GRIECO	100
GRIMALDI	100	NAPOLI	100
VALENTI	100	AMOROSO	100
PETRUCCI	100	DE MARTINO	100
MATTEI	100	IORIO	100
IZZO	100	LIGUORI	100
GENTILI	100	CALABRESE	100
ANTONELLI	100	DI MAIO	100
NATALE	100	BALDI	100
DE MARCO	100	NAPOLITANO	100
BERNARDINI	100	BORRELLI	100
BLASI	100	SESSA	100
BONANNI	100	MELE	100

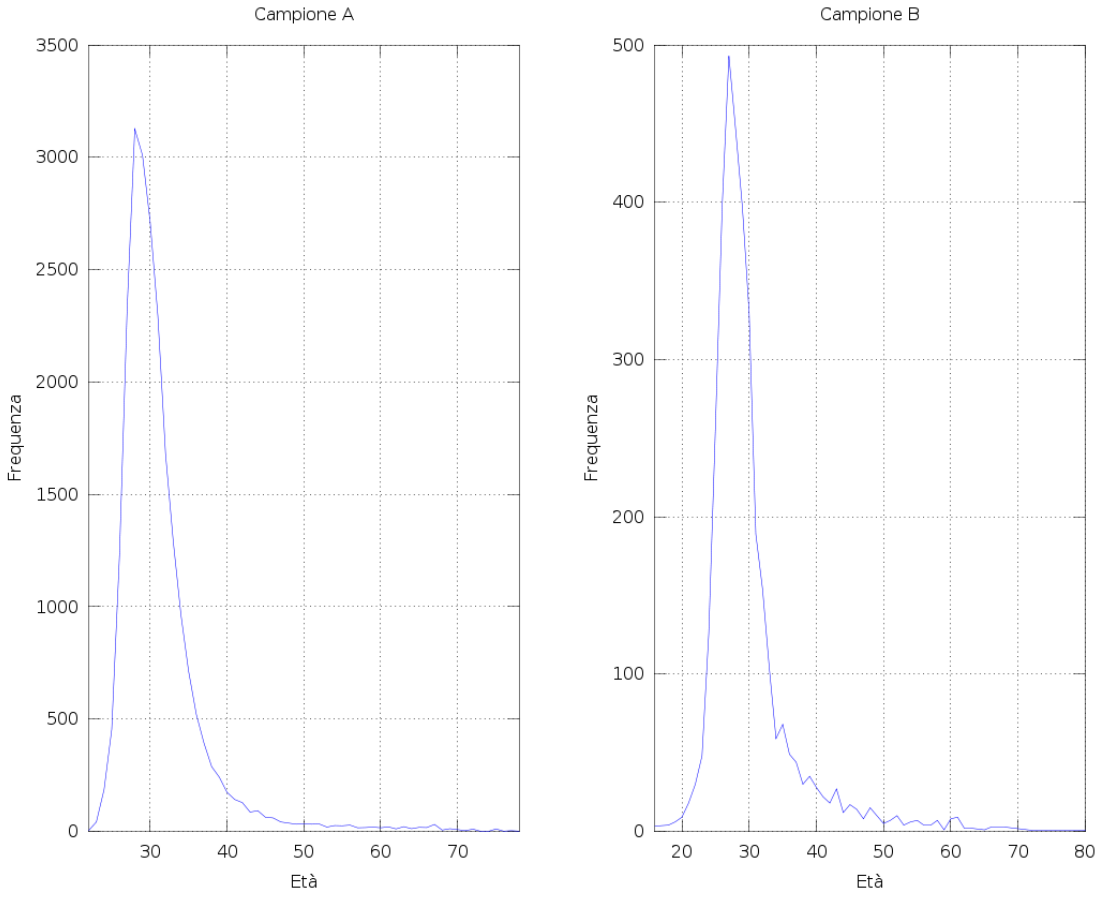
Osservando la Tabella 2 si può notare che solamente pochi cognomi mantengono la stessa posizione nella classifica rispetto al valore nazionale. Soprattutto nel campione B sono numerosi i campioni decisamente meno frequenti rispetto al dato di riferimento.
Ancora più evidente appare la presenza di cognomi non presenti nella classifica nazionale che costituiscono, sia per il campione A che per il campione B circa il 40% dei primi 100 cognomi.
Per quanto riguarda il campione A, trattandosi di una anagrafica molto localizzata, i risultati sono stati confrontati con l'elenco dei primi 20 cognomi presenti nella provincia di Roma, notando che molti dei cognomi non presenti nell'elenco nazionale dei primi 100 sono in realtà molto diffusi a livello provinciale. Questo giustifica il guadagno di posizione di cognomi come De Santis o De Angelis (4° e 5° nella provincia), o la presenza nei primi 100 del campione del cognome Proietti, secondo nella classifica provinciale.

Età al reclutamento

Le analisi successive riguardano l'età di inizio attività.
In questo caso la distribuzione attesa sarà di tipo normale. Dall'osservazione dei grafici emerge questo tipo di distribuzione anche se con una lunga coda a destra.
La distribuzione evidenziata sembra seguire la curva di Gompertz, un modello matematico nel quale il coefficiente di crescita è più lento nella fase iniziale e finale della curva; in particolare l'asintoto destro (rispetto al picco della curva), ha un approccio più graduale.

Per quanto riguarda il Campione A si è assunta la data di iscrizione all'Albo come data di inizio attività; tuttavia è evidente che l'iscrizione non comporta necessariamente l'avvio della professione di avvocato anche se ne è condizione necessaria.
Il campione B, relativamente ai dati oggetto di analisi, presenta numerose incongruenze dovute ad errori di inserimento nella base dati (ad esempio scambio delle due date, palese errore nell'anno, ecc.) e questo ha generato dati impossibili (età negative o superiori a mille). Nell'analisi pertanto si è provveduto a definire la scala, prendendo in considerazione età comprese tra 16 e 80 anni.

Considerando l'intero campione (maschi e femmine di qualsiasi età), i grafici risultanti sono i seguenti:

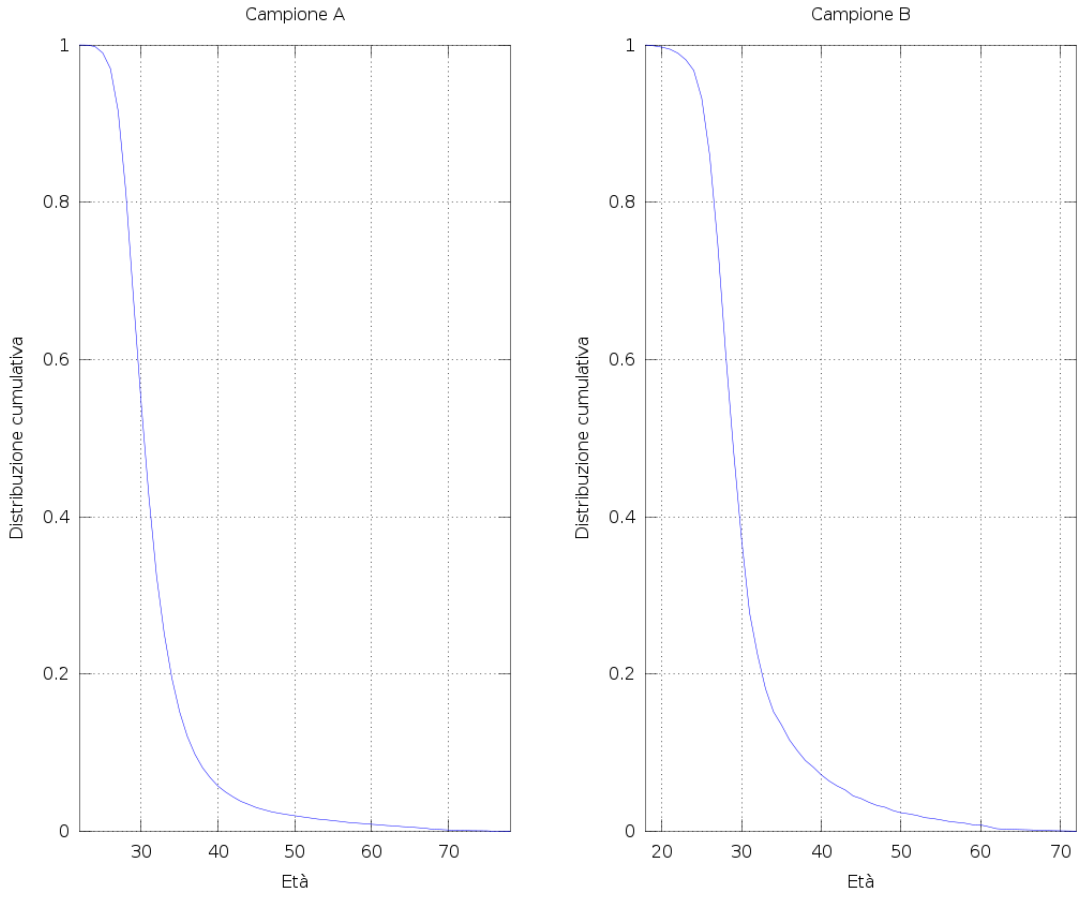


Le due distribuzioni appaiono decisamente simili, con un picco in corrispondenza dei 28 anni ed una lunga coda, piuttosto irregolare per il campione B.
Delle due distribuzioni è stata calcolata la distribuzione cumulativa definibile come

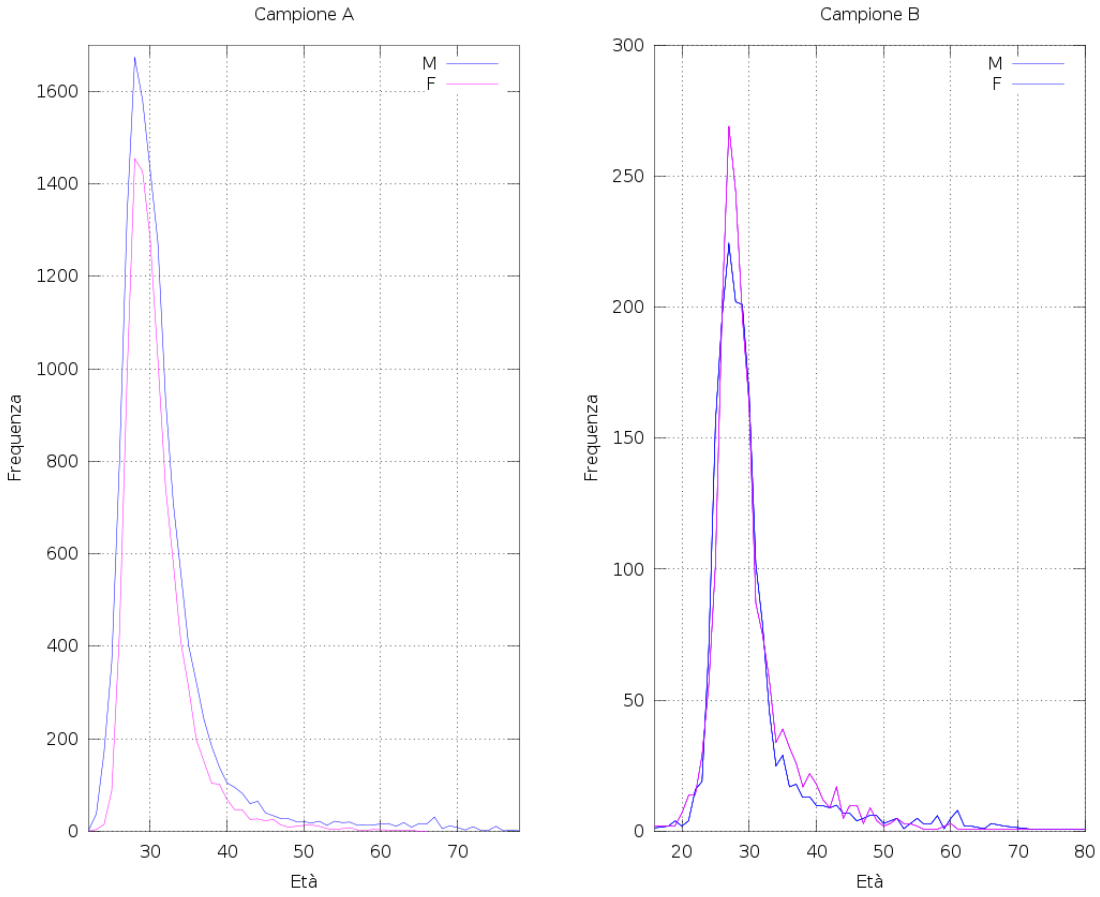
$$F(x)=P(X\geq x)$$

ed ottenuta sommando, per ogni valore di x, tutti i valori della distribuzione maggiori o uguali a x e dividendo per il totale dei valori. In questo modo si ottiene, per ogni valore di x, un valore compreso tra 0 e 1[3].
Rappresentata sul grafico, la curva evidenzia l'asimmetria della distribuzione e la lunghezza della coda destra.

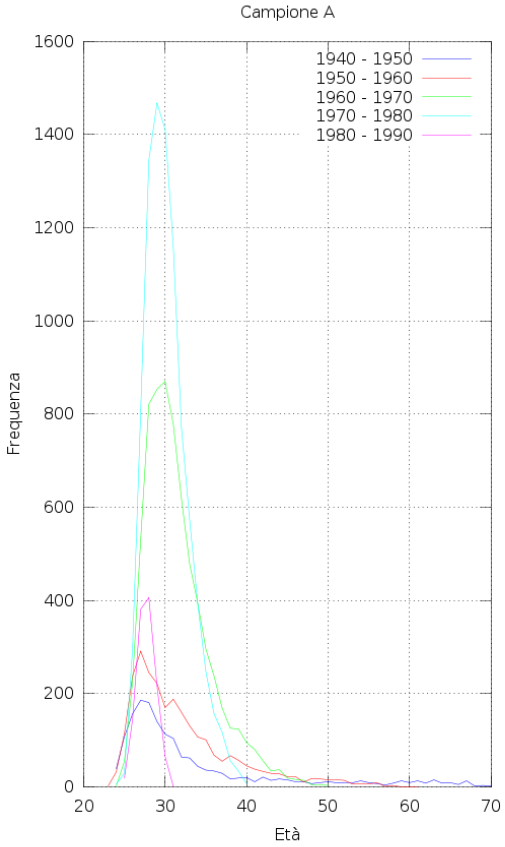
3 Il calcolo della distribuzione cumulativa è stato automatizzato mediante uno script in Python.

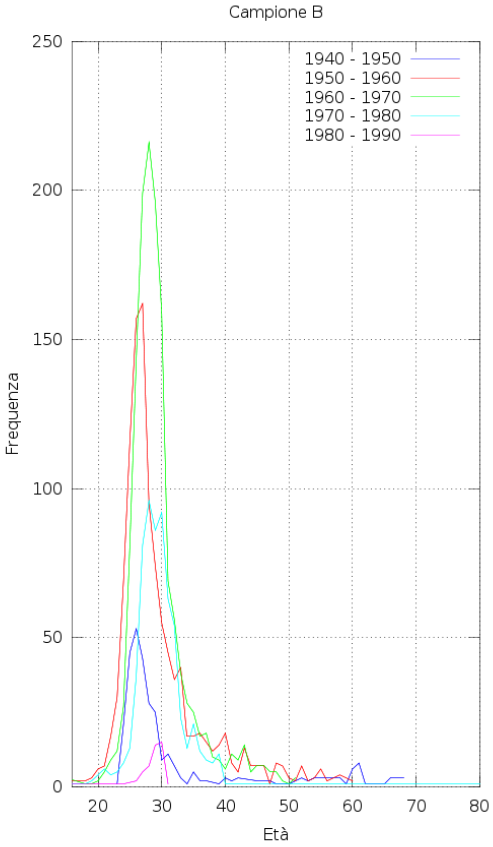


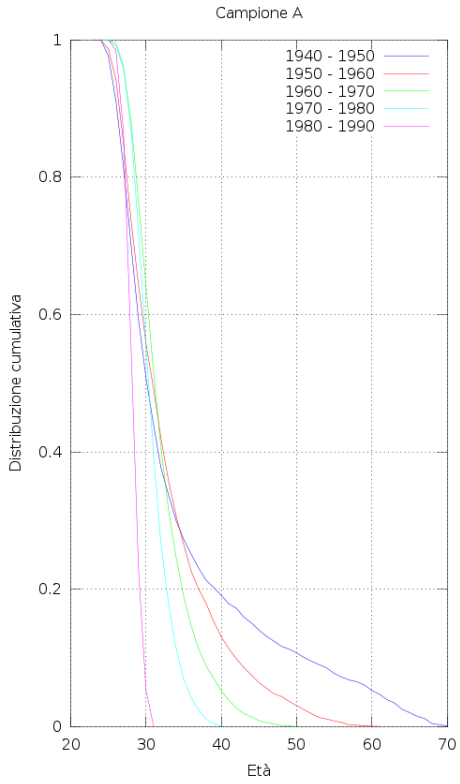
Ripetendo l'analisi sovrapponendo grafici sperati per i maschi (in blu) e per le femmine (in rosa), si nota che la distribuzione è pressoché la solita per i due gruppi, con il picco in corrispondenza dello stesso valore. Per quanto riguarda il campione A, la curva relativa alle femmine si presenta più stretta. Il numero delle donne iscritte all'Albo, inoltre, risulta inferiore a quello degli uomini. Al contrario, nel campione B, la curva delle femmine si sovrappone a quella dei maschi e le donne risultano numericamente superiori.

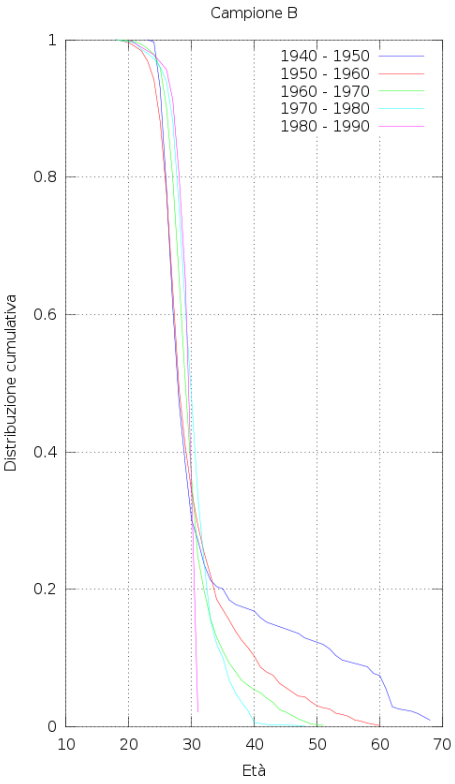


Infine, avendo estratto i dati per decennio di nascita, è possibile sovrapporre i dati in modo da confrontare in ottica temporale i vari valori. La scelta delle fasce di età può apparire arbitraria e riduttiva rispetto ad una analisi continuativa per anno di nascita. Tuttavia, non disponendo di dati rilevati in modo continuo nel tempo ma solo della situazione al momento dell'estrazione dei campioni, è possibile fornire solamente un'analisi della distribuzione attuale rispetto all'età dei soggetti correntemente in servizio. La divisione per decenni può quindi fornire informazioni interessanti su eventuali macro derive.
Nel campione A si nota un innalzamento dell'età di iscrizione all'albo per la generazione '60 – '70 mentre la tendenza sembra invertirsi per le generazioni successive. Nel campione B, al contrario, l'innalzamento dell'età media sembra essere progressivo.
Per quanto riguarda gli avvocati, l'età di iscrizione all'albo è legata prevalentemente alla durata degli studi e al superamento dell'esame pertanto sono da aspettarsi valori costanti rispetto al personale giudiziario per il quale, trattandosi di pubblico impiego, la deriva tenderà a seguire il trend nazionale dell'età media di inserimento nel mondo del lavoro.
Anche per queste distribuzioni è stata calcolata la distribuzione cumulativa, riportando sul medesimo grafico tutte le fasce di età
Il coefficiente di decrescita (ovvero la pendenza della curva) risulta essere maggiore nel campione B rispetto al campione A indicando che la possibilità di vincere un concorso nel pubblico impiego sembra ridursi con l'aumentare dell'età mentre le possibilità di superare l'esame per avvocato permangono più a lungo.


Campione A
1940 - 1950
1950 - 1960
1960 - 1970
1970 - 1980
1980 - 1990
Frequenza
Età
0
200
400
600
800
1000
1200
1400
1600
20
30
40
50
60
70


Campione B
1940 - 1950
1950 - 1960
1960 - 1970
1970 - 1980
1980 - 1990
Frequenza
Età
0
50
100
150
200
250
20
30
40
50
60
70
80


Campione A
1940 - 1950
1950 - 1960
1960 - 1970
1970 - 1980
1980 - 1990
Distribuzione cumulativa
Età
0
0.2
0.4
0.6
0.8
1
20
30
40
50
60
70


Campione B
1940 - 1950
1950 - 1960
1960 - 1970
1970 - 1980
1980 - 1990
Distribuzione cumulativa
Età
0
0.2
0.4
0.6
0.8
1
10
20
30
40
50
60
70

BIBLIOGRAFIA RAGIONATA

Una più che esaustiva bibliografia sullo studio della distribuzione dei cognomi è quella fornita da Paolo Rossi all'indirizzo http://www.df.unipi.it/~rossi/ a cui si rimanda, elencando invece in questa sede i testi che hanno fatto da supporto all'elaborazione del presente lavoro.

Per l'utilizzo del software Octave/Matlab sono risultati utili due manuali reperibili online:
P.J.G. Long, *Introduction to Octave*, Cambridge, 2005
http://www-mdp.eng.cam.ac.uk/web/CD/engapps/octave/octavetut.pdf
G. Arioli et al., *Laboratorio 1 - Introduzione a Matlab – Octave*, Politecnico di Milano, 2012
http://www1.mate.polimi.it/CN/MANI/lab/lab01/lab01.pdf

Sulla power law, il testo di riferimento è stato:
M.E.J. Newmann *Power laws. Pareto distributions and Zipf's law.* Contemporary Physics 46, 323 - 351 (2005)
Interessanti considerazioni sull'argomento, applicato alla teoria delle reti, si trovano in:
M. Buchanan, *Nexus*, Mondadori, Milano, 2003

Le analisi relative al reclutamento sono state elaborate avendo come riferimento:
P. Rossi, *Le dinamiche di reclutamento e di carriera dei fisici nel sistema universitario italiano*, in "Nuovo Saggiatore" 23, 3-4 (2007), p. 3

Informazioni di base sulla funzione di Gompertz sono state ricavate da
http://en.wikipedia.org/wiki/Gompertz_function
mentre sulla distribuzione cumulativa il riferimento è stato
http://en.wikipedia.org/wiki/Cumulative_distribution_function