The Truncated Poisson Models with covariates to estimate the size of drug users' population at risk of cautioning for personal possession

Flavia Mascioli[1] Carla Rossi[2] Daria Scacciatelli[2]

[1]Department of Mathematics
University of Rome "La Sapienza"

[2]Centre for Biostatistics and Bioinformatics
University of Rome Tor Vergata

Illecit Drug Market Workshop, 2009

Outline

Introduction

- **Study Objective**
  Estimate the number of drug users' at risk of cautioning for personal possession in Italy, in the year 2007.
  Show the robustness of our results by a bootstrap simulation study.
- **Data Source**
  Real data, provided by the Italian Ministry of Interiors in which individuals are divided by sex, age class, geographical area and number of cautionings
  Simulated data, where the parameters were chosen according to the real data.
- **Methods**
  Truncated Poisson model with covariates, Horvitz-Thompson, Zelterman and Chao estimators and the Bootstrap method, are considered.

Formulation of the problem

- Let $n_1, n_2, ..., n_m$ be the frequencies of individuals identified $1, 2, ..., m$ times and let $p_1, p_2, ..., p_m$ be the associated probabilities.
- The frequency n_0, the unobserved number of individuals who are identified zero times, is missing and constitutes the target of the inference
- A population has N units n of which are identified by some mechanism. If the probability to identify an unit is $(1 - p_0)$:

$$N = Np_0 + (1 - p_0)N = unobserved + observed \ \ cases = Np_0 + n$$

The zero-truncated Poisson Model

If p_0 is unknown, as happen in most of the real cases, a conventional approach assumes that the frequencies arise from a Zero-Truncated Poisson Model:

$$P_j = P_j(\lambda) = exp(-\lambda)\frac{\lambda^j}{j!} \quad j = 1, ...,$$

where the parameter λ can be estimated by different methods (i.e. Maximum likelihood approach).

Underlying Assumptions

- **Closed population** (i.e. no in-migration or out-migration in the time period studied)
- **Homogeneous population** (i.e. no subgroups with markedly different probabilities to be observed and re-observed)
- **Constant probability of being observed** (i.e. the probability of being re-observed should not be influenced by the experience of a previous visit)

Solutions:

- Considering a short time study period
- Introducing observable covariates
- Considering a short time study period and stratifying by geographical areas.

Population size Estimators

1. **Horvitz-Thompson** estimator (1952)

$$\widehat{N}_{HT} = \frac{n}{(1 - \widehat{p}_0)}$$

   this model is not flexible enough in capturing population heterogeneity, and will generally underestimate the population size

2. lower bound estimator by **Chao** (1987, 1989)

$$\widehat{N}_C = n + \frac{n_1^2}{2n_2}$$

   based on a mixture of Poisson distributions

3. robust estimator of **Zelterman** (1988)

$$\widehat{N}_Z = \frac{n}{(1 - e^{-2\frac{n_2}{n_1}})}$$

   based only on the lower frequencies n_1 and n_2

Confidence intervals for $\widehat{N}$

- A natural approach to construct the confidence intervals for $\widehat{N}$ assumes that $\widehat{N}$ is asymptotically normal:

$$\widehat{N} \pm 1.96\sqrt{var(\widehat{N})}$$

.

- Chao (1989) proposed to reparametrize the model using: $log(\widehat{N} - n)$, where n is the number of different individuals identified.
- For Zelterman and Chao estimators, we consider the new variances formulation proposed by Böhning (2008).

Real Data

- Datasets provided by the Italian Ministry of Interiors: individuals are identified for the first time in 2007 and are divided by sex, age class, geographical area and number (one or more than once) of cautionings.
- In 2008, Mascioli and Rossi compare the three estimators with covariates sex and age class.

Both estimators produce almost the same estimates of N, for each age class and for the total population. As expected, when λ is small, Zelterman estimates are slightly greater than Chao estimates (Böhning and Brittain, 2007).

Males	*Estimates*			*Estimates*	
Age Classes	$\widehat{\lambda}$	$\widehat{N}_C$	95%*CI*	$\widehat{N}_Z$	95%*CI*
< 15	0.076	5706	3599-9213	5701	3565-9290
15-17	0.066	31179	24543-39767	31188	24455-39939
18-19	0.069	59742	50573-70714	59755	50440-70937
20-24	0.067	111588	98334-126768	111629	98164-127086
25-29	0.043	96439	77994-119480	96505	77873-119834
30-34	0.045	53480	40800-70339	53508	40700-70590
35-39	0.052	31345	23201- 42556	31350	23116-42733
> 39	0.051	25026	17700-35611	25036	17631-35783
Total (No covariate)	0.059	398857	369676-430492	399012	369407-431145

Females	*Estimates*			*Estimates*	
Age Classes	$\widehat{\lambda}$	$\widehat{N}_C$	95%*CI*	$\widehat{N}_Z$	95%*CI*
< 15	0.0500	801	182-4104	800	180-4165
15-17	0.0430	3268	1201-9359	3267	1190-9451
18-19	0.0370	7401	3294-17088	7399	3271-17208
20-24	0.0300	21719	12007 -39743	21716	11954-39915
25-29	0.0310	12484	5898-26941	12482	5865-27089
30-34	0.0290	7282	2632-20850	7281	2616-20987
35-39	0.0430	3362	1235-9628	3361	1224-9722
> 39	0.0190	5833	1209-29861	5832	1203-30024
Total (No covariate)	0.033	61725	43761-87409	61745	43656-87682

For the highlighted areas the capture probability is smaller than in the other areas and the corresponding confidence intervals are larger.

Areas		*Estimates*		*Estimates*	
	$\widehat{\lambda}$	$\widehat{N}_C$	95%*CI*	$\widehat{N}_Z$	95%*CI*
Basilicata	0.083	3087	1722-5714	3090	1704-5789
Calabria	0.057	12736	8364-19606	12731	8313-19719
Campania	0.117	11696	9390-14660	11693	9327-14754
EmiliaRomagna	0.059	35063	27437-44980	35040	27319-45120
FriuliVeneziaGiulia	0.053	4499	2139-9778	4500	2121-9873
Lazio	0.036	54901	39210-77189	54938	39123-77470
Liguria	0.051	23760	16789-33845	23753	16712-33989
Lombardia	0.050	66949	54142-82982	66934	53986-83190
Marche	0.048	15113	9551-24175	15112	9500-24309
Abruzzo + Molise	0.061	10605	6883-16541	10603	6839-16650
Piemonte + Vald' Aosta	0.051	50010	39156-64076	50009	39035-64276
Puglia	0.063	26622	20274-35129	26629	20195-35291
Sardegna	0.035	25293	15459-41755	25297	15397-41938
Sicilia	0.079	40013	33420-48034	40037	33324-48235
Trentino − AltoAdige	0.051	6466	3359-12749	6467	3334-12855
Toscana	0.044	45111	33576-60853	45101	33461-61040
Umbria	0.020	14202	5166-40008	14196	5140-40183
Veneto	0.050	37247	28095-49590	37238	27989-49759
Total (No covariate)	0.057	434490	403930-467510	434230	403260-467730

Both Chao and Zelterman estimates are greater than Horvitz-Thompson estimates.

Areas	*Data*			*Estimates*				
Age Classes	n	n_1	$n_{>1}$	n_2	$\widehat{\lambda}$	$\widehat{N}_C$	$\widehat{N}_Z$	$\widehat{N}_{HT}$
Basilicata	243	233	10	10	0.083	3087	3090	3035
Calabria	712	692	20	20	0.057	12736	12731	12913
Campania	1294	1220	74	72	0.117	11696	11693	11753
EmiliaRomagna	2054	1994	60	60	0.059	35063	35040	35849
FriuliVeneziaGiulia	230	224	6	6	0.053	4499	4500	4486
Lazio	1885	1851	34	32	0.036	54901	54938	52885
Liguria	1190	1160	30	30	0.051	23760	23753	24002
Lombardia	3301	3219	82	81	0.050	66949	66934	67552
Marche	712	695	17	17	0.048	15113	15112	15149
Abruzzo + Molise	630	611	19	19	0.061	10605	10603	10657
Piemonte + Vald' Aosta	2471	2409	62	61	0.051	50010	50009	50071
Puglia	1608	1558	50	49	0.063	26622	26629	26398
Sardegna	863	848	15	15	0.035	25293	25297	25115
Sicilia	2985	2869	116	111	0.079	40013	40037	39414
Trentino − AltoAdige	318	310	8	8	0.051	6466	6467	6427
Toscana	1943	1901	42	42	0.044	45111	45101	45596
Umbria	297	294	3	3	0.020	14202	14196	14801
Veneto	1845	1799	46	46	0.050	37247	37238	37620
Total (No covariate)	24581	23887	694	696	0.057	434490	434230	443590

Simulated Data

covariates: sex, age class, geographical area

- For every covariate we simulate, by a Binomial distribution, a population of 10000 units and we make a Bootstrap with 100 replications, to estimate the variance of $\widehat{N}$
- The observed population size n and capture-recapture probability were chosen according to the real data.

Comparison between the results obtained both on real data and on simulated data. For males population, the results are superimposable.

Males	*Estimates*		*Estimates*	
Age Classes	$\widehat{N}_C$	95%*CI*	$\widehat{N}_Z$	95%*CI*
< 15	5706	3599-9213	5701	3565-9290
15-17	31179	24543-39767	31188	24455-39939
18-19	59742	50573-70714	59755	50440-70937
20-24	111588	98334-126768	111629	98164-127086
25-29	96439	77994-119480	96505	77873-119834
30-34	53480	40800-70339	53508	40700-70590
35-39	31345	23201- 42556	31350	23116-42733
> 39	25026	17700-35611	25036	17631-35783
Total (No covariate)	398857	369676-430492	399012	369407-431145

Males	*Bootstrap Estimates*		*Bootstrap Estimates*	
Age Classes	$\widehat{N}_C^B(STD)$	95%*CI*	$\widehat{N}_Z^B(STD)$	95%*CI*
< 15	6305(1922)	3574-11405	6307(1922)	3576-11408
15-17	31924(3964)	25106-40752	31927(3960)	25115-40744
18-198	61178(4738)	52619-71242	61188(4736)	52631-7124
20-24	110790(6809)	98281-125016	110793(6816)	98270-125036
25-29	94258(10612)	75738-117558	94274(10616)	75748-117584
30-34	54068(8901)	39348-74637	54075(8915)	39336-74682
35-39	32227(4689)	24326-42873	32232(4691)	24327-42883
> 39	26158(6213)	16637-41561	26157(6219)	16629-41578
Total (No covariate)	394016(16177)	363638-427100	393975(16227)	363507-427165

For females population, the results are less significant because the recapture probability is very low.

Females	*Data*		
Age Class	n	n_1	$n_{>1}$
< 15	40	39	1
15-17	140	137	3
18-19	272	267	5
20-24	659	649	10
25-29	387	381	6
30-34	209	206	3
35-39	142	139	3
> 39	108	107	1
Total (No covariate)	1957	1925	32

The results obtained with the bootstrap simulation study are comparable with the ones obtained on the real data;

Areas		*Estimates*	*Bootstrap Estimates*	
Chao	$\widehat{N}_C$	95%*CI*	$\widehat{N}_C^B(STD)$	95%*CI*
Basilicata	3087	1722-5714	3197 (1151)	1656-6415
Calabria	12736	8364-19606	14064(3774)	8466-23703
Campania	11696	9390-14660	12052(1490)	9505-15389
EmiliaRomagna	35063	27437-44980	35975(5006)	27494-47283
FriuliVeneziaGiulia	4499	2139-9778	5240 (2211)	2422-11682
Lazio	54901	39210-77189	56565(14468)	34725-92930
Liguria	23760	16789-33845	25166 (5444)	16640-38399
Lombardia	66949	54142-82982	69463 (7092)	56958-84882
Marche	15113	9551-24175	17018 (4953)	9821-29901
Abruzzo + Molise	10605	6883-16541	11418 (2697)	7288-18109
Piemonte + Vald'Aosta	50010	39156-64076	50750 (7375)	38320-67490
Puglia	26622	20274-35129	27314 (4047)	20524-36541
Sardegna	25293	15459-41755	30442 (13678)	13344-70960
Sicilia	40013	33420-48034	40249 (3579)	33868-47948
Trentino − AltoAdige	6466	3359-12749	7175 (3587)	2933-18299
Toscana	45111	33576-60853	48038 (7930)	34925-66364
Umbria	14202	5166-40008	39253 (6225)	28902-53563
Veneto	37247	28095-49590	38707 (5441)	29490-50997
Total (No covariate)	434490	403930-467510	443230 (18125)	409190-480290

Areas		*Estimates*	*Bootstrap Estimates*	
Zelterman	$\widehat{N}_Z$	95%*CI*	$\widehat{N}_Z^B(STD)$	95%*CI*
Basilicata	3090	1704-5789	3198(1152)	1657-6419
Calabria	12731	8313-19719	14066 (3778)	8463-23718
Campania	11693	9327-14754	12055 (1494)	9503-15401
EmiliaRomagna	35040	27319-45120	35973 (5007)	27490-47283
FriuliVeneziaGiulia	4500	2121-9873	5243 (2212)	2423-11687
Lazio	54938	39123-77470	56582 (14483)	34722-92992
Liguria	23753	16712-33989	25167 (5449)	16634-38414
Lombardia	66934	53986-83190	69466 (7104)	56942-84913
Marche	15112	9500-24309	17024 (4960)	9819-29930
Abruzzo + Molise	10603	6839-16650	11419 (2698)	7288-18115
Piemonte + Vald'Aosta	50009	39035-64276	50748 (7382)	38308-67508
Puglia	26629	20195-35291	27315 (4048)	20524-36545
Sardegna	25297	15397-41938	30450 (13689)	13343-71010
Sicilia	40037	33324-48235	40270 (3576)	33895-47961
Trentino − AltoAdige	6467	3334-12855	7177 (3588)	2934-18305
Toscana	45101	33461-61040	48051 (7941)	34922-66406
Umbria	14196	5140-40183	39255 (6234)	28892-53589
Veneto	37238	27989-49759	38708 (5452)	29476-51026
Total (No covariate)	434230	403260-467730	443190(18213)	408990-480440

Comparison between the three estimators considering the model with or without covariates. Considering covariates allow to obtain better results

Real Data	$\widehat{N}_C$	$\widehat{N}_{HT}$	$\widehat{N}_Z$
Males (all age covariate)	414504	409692	414672
Males (no covariate)	398857	394898	399012
Females (all age covariate)	63110	61391	63136
Females (no covariate)	66298	66394	60487

Simulate Data	$\widehat{N}_C$	$\widehat{N}_{HT}$	$\widehat{N}_Z$
Males (all age covariate)	416907	415048	416952
Males (no covariate)	394016	395404	393975
Females (all age covariate)	75102	72272	75127
Females (no covariate)	64215	62101	64248

Real Data	$\widehat{N}_C$	$\widehat{N}_{HT}$	$\widehat{N}_Z$
Areas (all area covariate)	483374	483722	483367
Areas (no covariate)	434490	443590	434230

Simulate Data	$\widehat{N}_C$	$\widehat{N}_{HT}$	$\widehat{N}_Z$
Areas (all area covariate)	532086	527670	532167
Areas (no covariate)	443230	444680	443190

Comparison between the three estimators considering the model with or without covariates. Considering covariates allow to obtain better results

Real Data	$\widehat{N}_C$	$\widehat{N}_{HT}$	$\widehat{N}_Z$
Males (all age covariate)	414504	409692	414672
Males (no covariate)	398857	394898	399012
Females (all age covariate)	63110	61391	63136
Females (no covariate)	66298	66394	60487

Simulate Data	$\widehat{N}_C$	$\widehat{N}_{HT}$	$\widehat{N}_Z$
Males (all age covariate)	416907	415048	416952
Males (no covariate)	394016	395404	393975
Females (all age covariate)	75102	72272	75127
Females (no covariate)	64215	62101	64248

Real Data	$\widehat{N}_C$	$\widehat{N}_{HT}$	$\widehat{N}_Z$
Areas (all area covariate)	483374	483722	483367
Areas (no covariate)	434490	443590	434230

Simulate Data	$\widehat{N}_C$	$\widehat{N}_{HT}$	$\widehat{N}_Z$
Areas (all area covariate)	532086	527670	532167
Areas (no covariate)	443230	444680	443190

Conclusion and Future Work

Conclusion

- Considering covariates in the model produces more accurate estimates
- The bootstrap simulation study confirm the coherence of the results obtained with real data (with the exception of females population)

Future Work

- Analysis of differences between large and small areas (several results have been already obtained)
- Inclusion of more covariates in the model, socioeconomic background and behavior
- Use more sophisticated simulation schemes

References

- Böhning, D. (2008). A Simple Variance Formula for Population Size Estimators by Conditioning. Statistical Methodology, 5, 410-423
- Böhning, D. and Del Rio Vilas, V. (2008). Estimating the hidden number of scrapie affected holdings in Great Britain using a simple, truncated count model allowing for heterogeneity. Journal of Agricultural, Biological, and Environmental Statistics, 13, 1-22.
- Chao, A. (1987). Estimating the population size for capture-recapture data with unequal catchability. Biometrics, 43, 783-791.
- Chao, A. (1989). Estimating population size for sparse data in capture-recapture experiments. Biometrics, 45, 427-438.
- Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. Journal of the American Statistical Association, 47, 663-685.
- Mascioli F. and Rossi C.(2009). Capture-recapture methods to estimate prevalence indicators used in the evaluation of drug policies. 3rd Annual Conference of the International Society for the Study of Drug Policy, (Vienna, 2-3 March 2009).
- Rossi C. and Ricci R. (2009). Modelling and estimating illicit drug market as a tool to evaluate drug policy: the case of Italy. 3rd Annual Conference of the International Society for the Study of Drug Policy, (Vienna, 2-3 March 2009).
- Scalia Tomba GP., Rossi C., Taylor C., Klempova D., Wiessing L. (2008). Guidelines for Estimating the Incidence of Problem Drug Use. EMCDDA, Lisbon.
- Zelterman, D. (1988). Robust estimation in truncated discrete distributions with application to capture-recapture experiments. Journal of Statistical Planning and Inference, 18, 225-237.