I dati imputati: colpevoli o innocenti?

Massimiliano Pastore

PSICOSTAT meeting:
18 Dicembre 2020

Outline

Introduzione

Data imputation: quello che ho capito

- In statistics, **imputation** is the process of replacing missing data with substituted values.
- By far, the most common means of dealing with missing data is **listwise deletion** which is when all cases with a missing value are deleted.
- **Single imputation** (e.g. mean substitution or regression) does not take into account the uncertainty in the imputations.
- **Multiple Imputation** (Rubin, 1987) has become a generally accepted way to handle statistical analysis of incomplete data (Koller-Meinfelder, 2009).

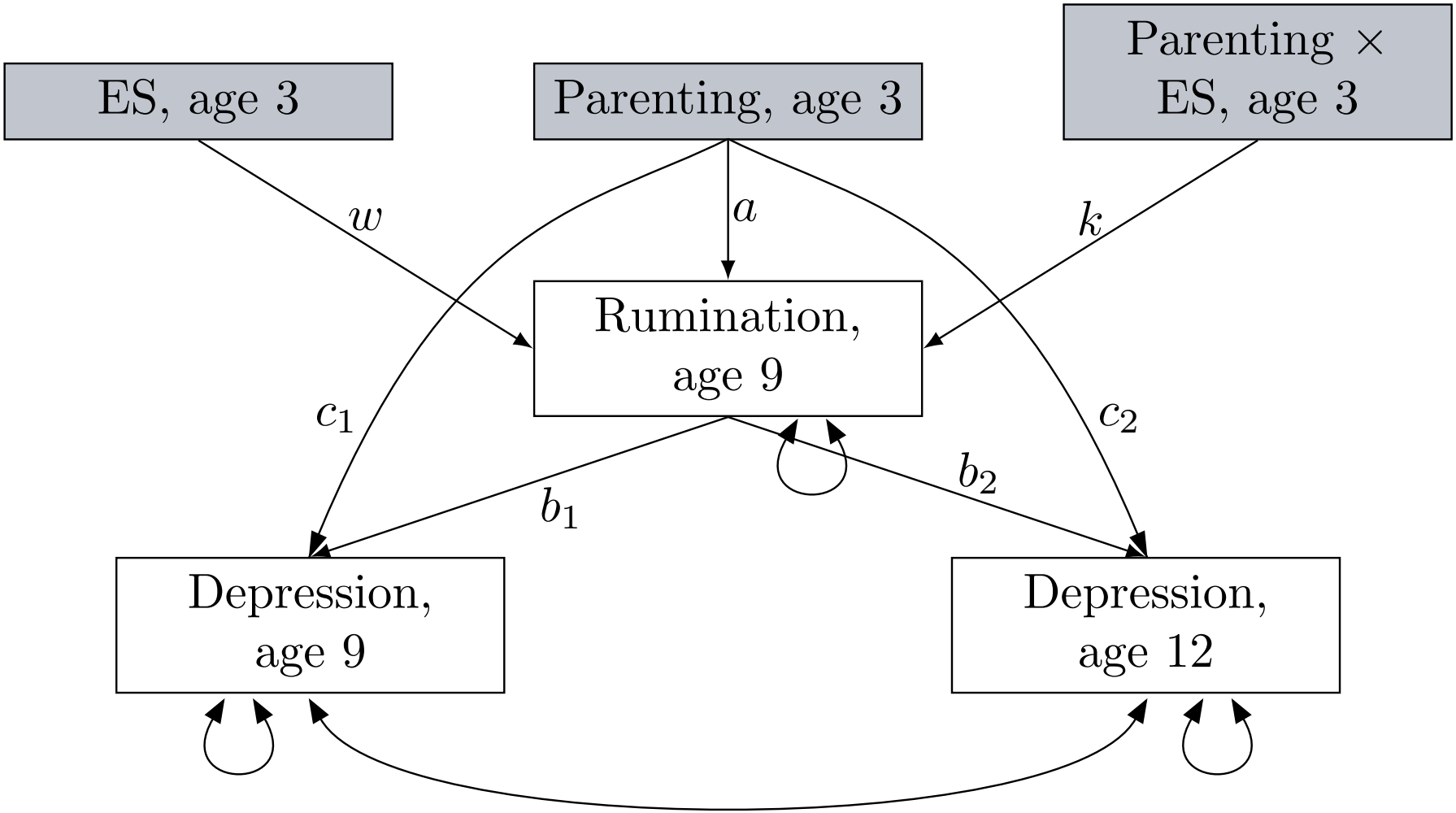
---

https://en.wikipedia.org/wiki/Imputation_(statistics)

Rubin, D.B. (1987). *Multiple imputation for survey nonresponse.* New York: Wiley.

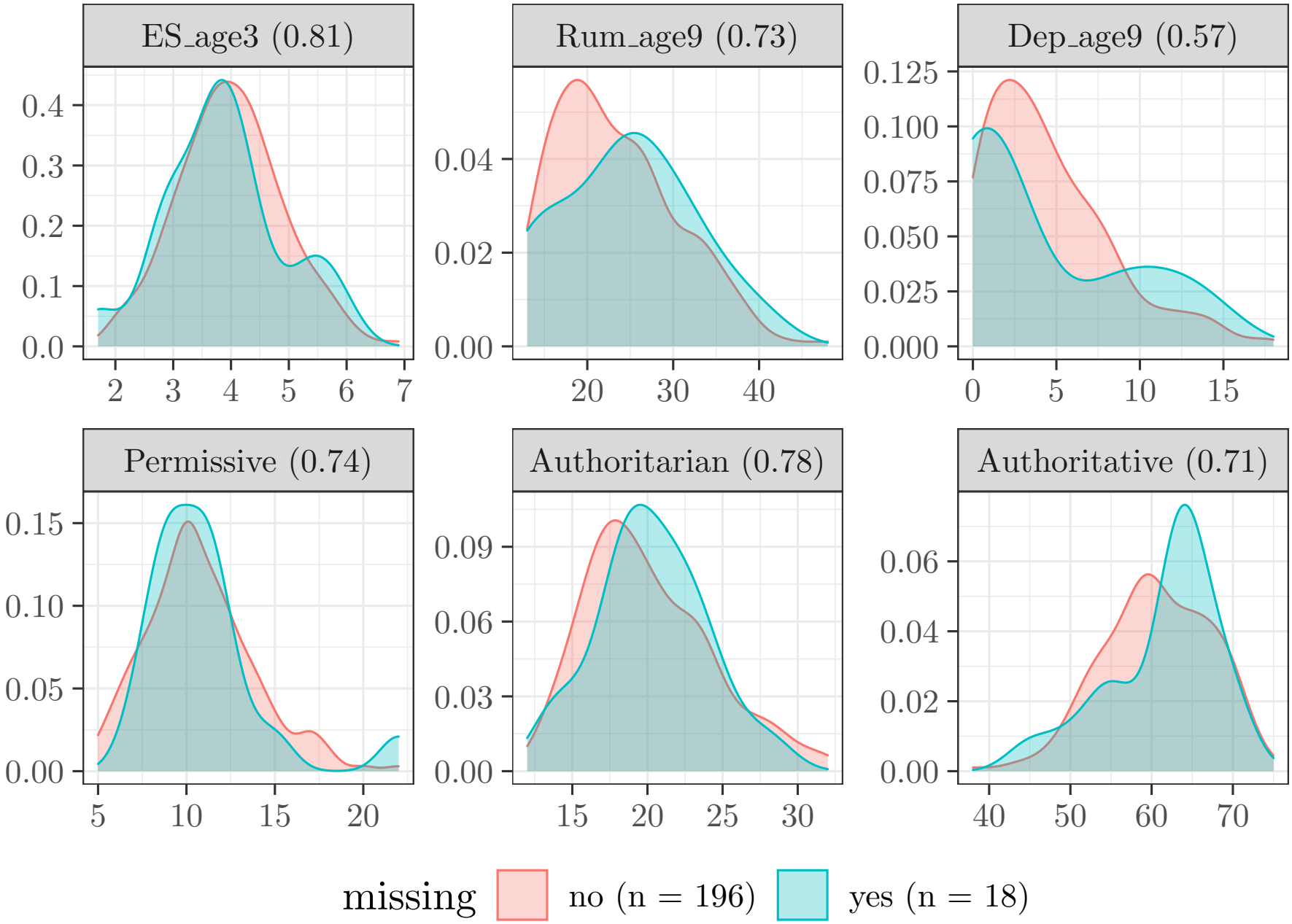
Koller-Meinfelder, F. (2009). *Analysis of Incomplete Survey Data-Multiple Imputation via Bayesian Bootstrap Predictive Mean Matching* (Unpublished doctoral dissertation).

Un esempio pratico


ES, age 3
Parenting, age 3
Parenting ×
ES, age 3
w
a
k
Rumination,
age 9
c1
c2
b1
b2
Depression,
age 9
Depression,
age 12

- Il campione si compone di 214 soggetti.
- Per 18 di essi mancano i dati nella variabile *Depression, age 12*.

	vars	n	mean	sd	min	max	range	se
ES_age3	1	214	3.99	0.91	2	7	5	0.06
Rum_age9	2	214	23.50	7.08	13	48	35	0.48
Dep_age9	3	214	4.50	3.87	0	18	18	0.26
Dep_age12	4	196	4.55	5.39	0	28	28	0.39
Permissive	5	214	10.77	3.06	5	22	17	0.21
Authoritarian	6	214	20.09	4.13	12	32	20	0.28
Authoritative	7	214	60.75	6.65	38	75	37	0.45


ES_age3 (0.81)
Rum_age9 (0.73)
Dep_age9 (0.57)
Permissive (0.74)
Authoritarian (0.78)
Authoritative (0.71)
missing
no (n = 196)
yes (n = 18)

```
Model Info:
 function:     stan_glm
 family:       binomial [logit]
 formula:      is.na ~ ES_age3 + Rum_age9 + Dep_age9 +
   Permissive + Authoritarian + Authoritative
 algorithm:    sampling
 sample:       4000 (posterior sample size)
 priors:       see help('prior_summary')
 observations: 214
 predictors:   7

Estimates:
                mean   sd    5%    95%
(Intercept)   -3.40   3.24 -8.75  2.03
ES_age3       -0.25   0.29 -0.73  0.23
Rum_age9       0.04   0.04 -0.03  0.10
Dep_age9      -0.07   0.08 -0.20  0.06
Permissive     0.02   0.09 -0.13  0.16
Authoritarian  0.00   0.07 -0.11  0.11
Authoritative  0.02   0.04 -0.05  0.08
```

Data Imputation

- Per l'imputazione dei dati mancanti abbiamo utilizzato una procedura di *Multiple Imputation through Bayesian Bootstrap Predictive Mean Matching* grazie al pacchetto `BaBooN` (Meinfelder & Schnapp, 2015).
- In pratica basta scrivere:

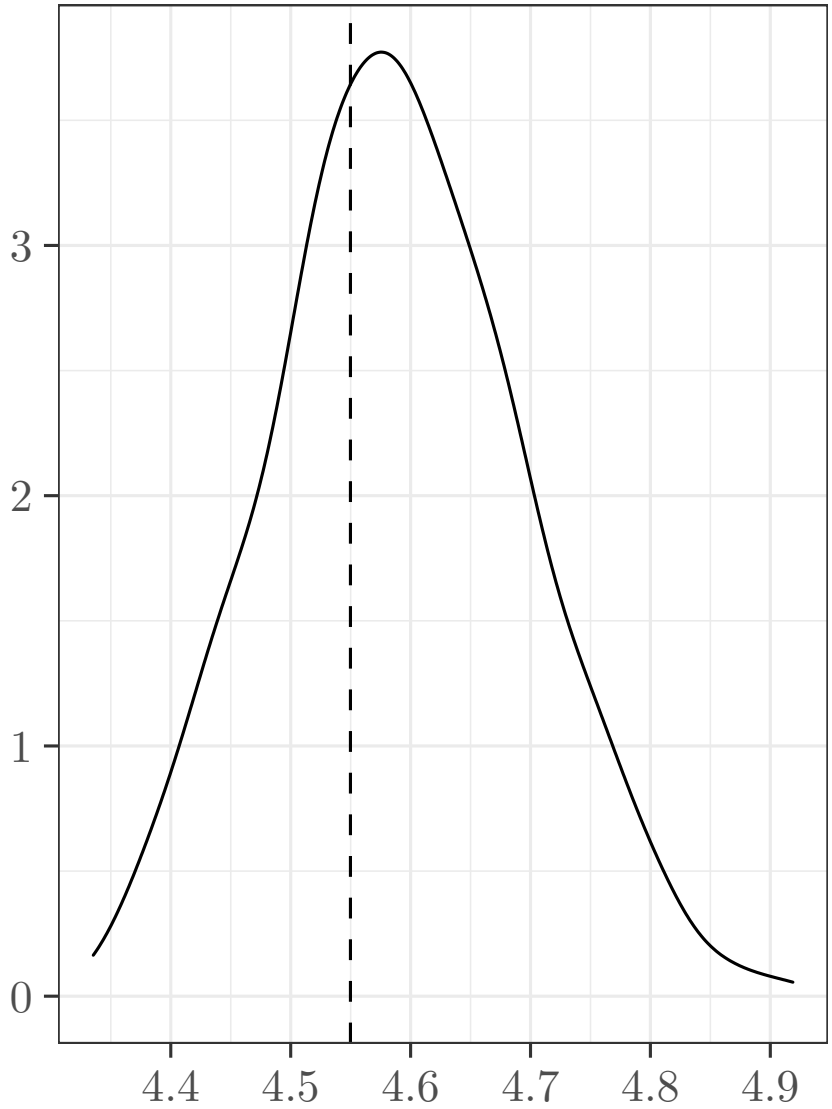
```
> BBPMM( data, M = M )
```

  in cui `data` è il dataset contenente la variabile contenente i missing e le variabili coinvolte nel processo e `M` il numero di imputazioni desiderate.
- Per valutare il funzionamento della procedura abbiamo ripetuto l'imputazione per 500 volte, ricalcolato la media della variabile `Dep_age12` e confrontato le distribuzioni dei valori ottenuti con quelli osservati.

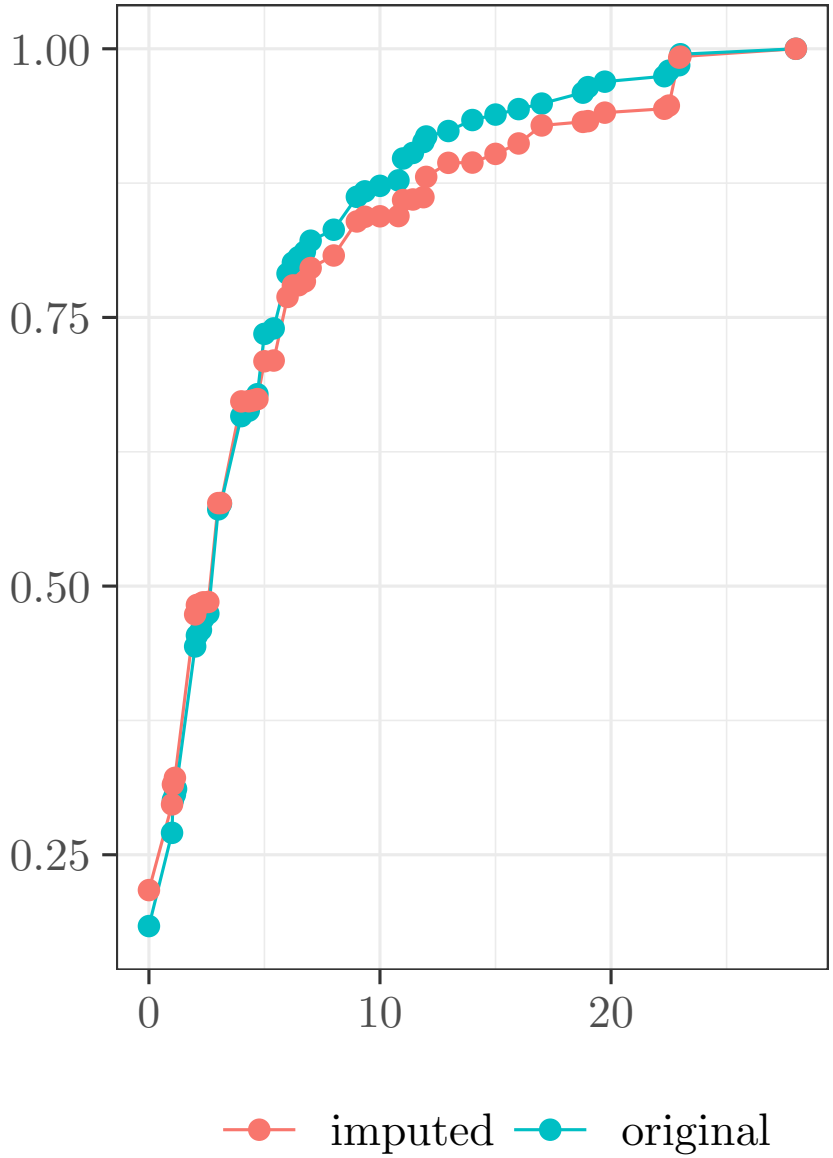
Meinfelder, F., & Schnapp, T. (2015). `BaBooN`: *Bayesian Bootstrap Predictive Mean Matching - Multiple and Single Imputation for Discrete Data*. Retrieved from `https://CRAN.R-project.org/package=BaBooN` (R package version 0.2-0).

Check on imputed data

[A] means



[B] cumulative



Sensitivity analysis

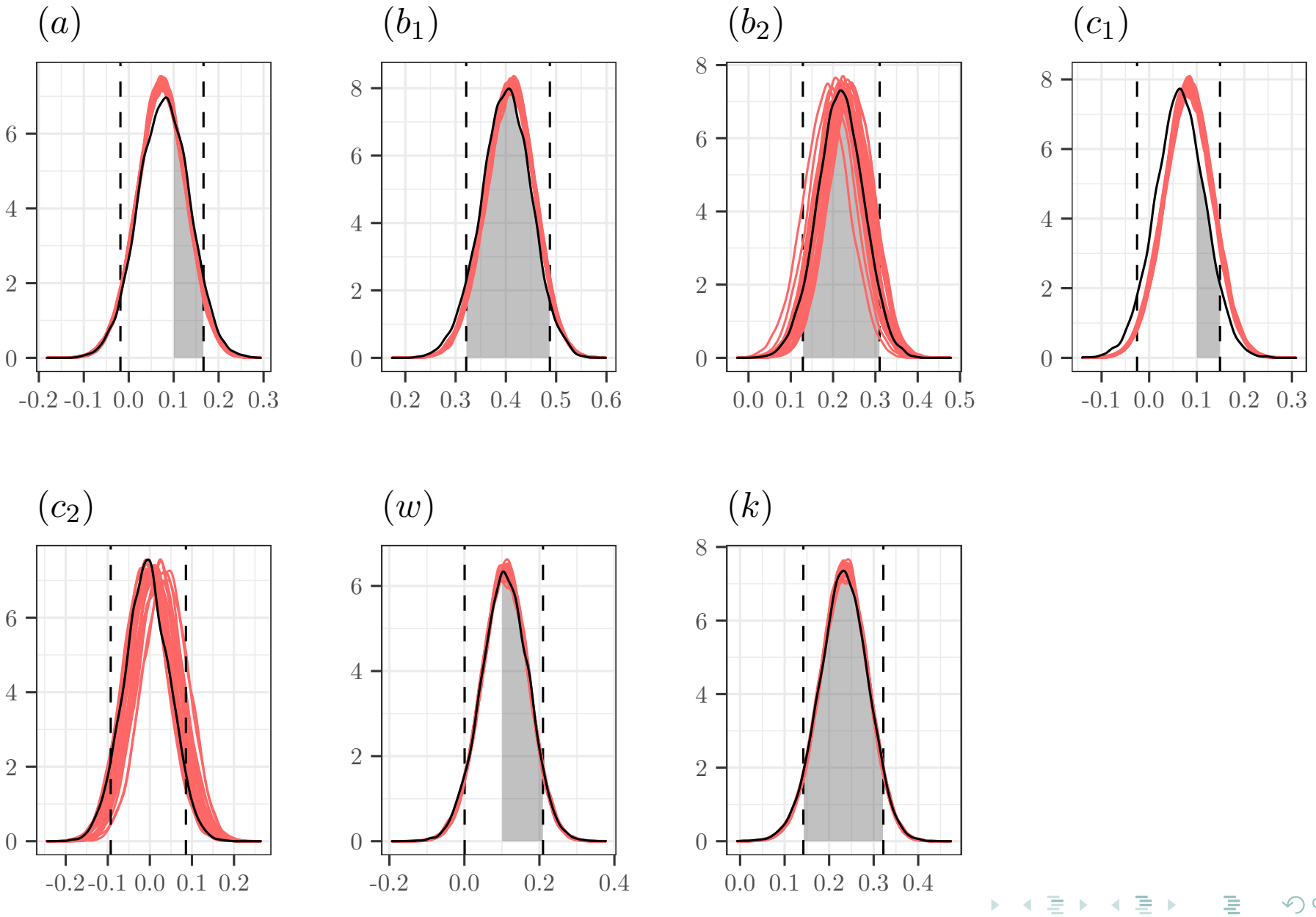
- A questo punto abbiamo selezionato 25 imputazioni e ristimato per altrettante volte le distribuzioni a posteriori dei parametri dei tre modelli con le stesse prior informative usate con i dati effettivi:

$$
\begin{aligned}
a &\sim \text{Normal}(0.1, 0.1) \\
b_1 &\sim \text{Normal}(0.5, 0.1) \\
b_2 &\sim \text{Normal}(0.35, 0.1) \\
c_1 &\sim \text{Normal}(\pm 0.1, 0.1) \\
c_2 &\sim \text{Normal}(\pm 0.05, 0.1) \\
w &\sim \text{Normal}(0, 0.2) \\
k &\sim \text{Normal}(0.3, 0.1)
\end{aligned}
$$

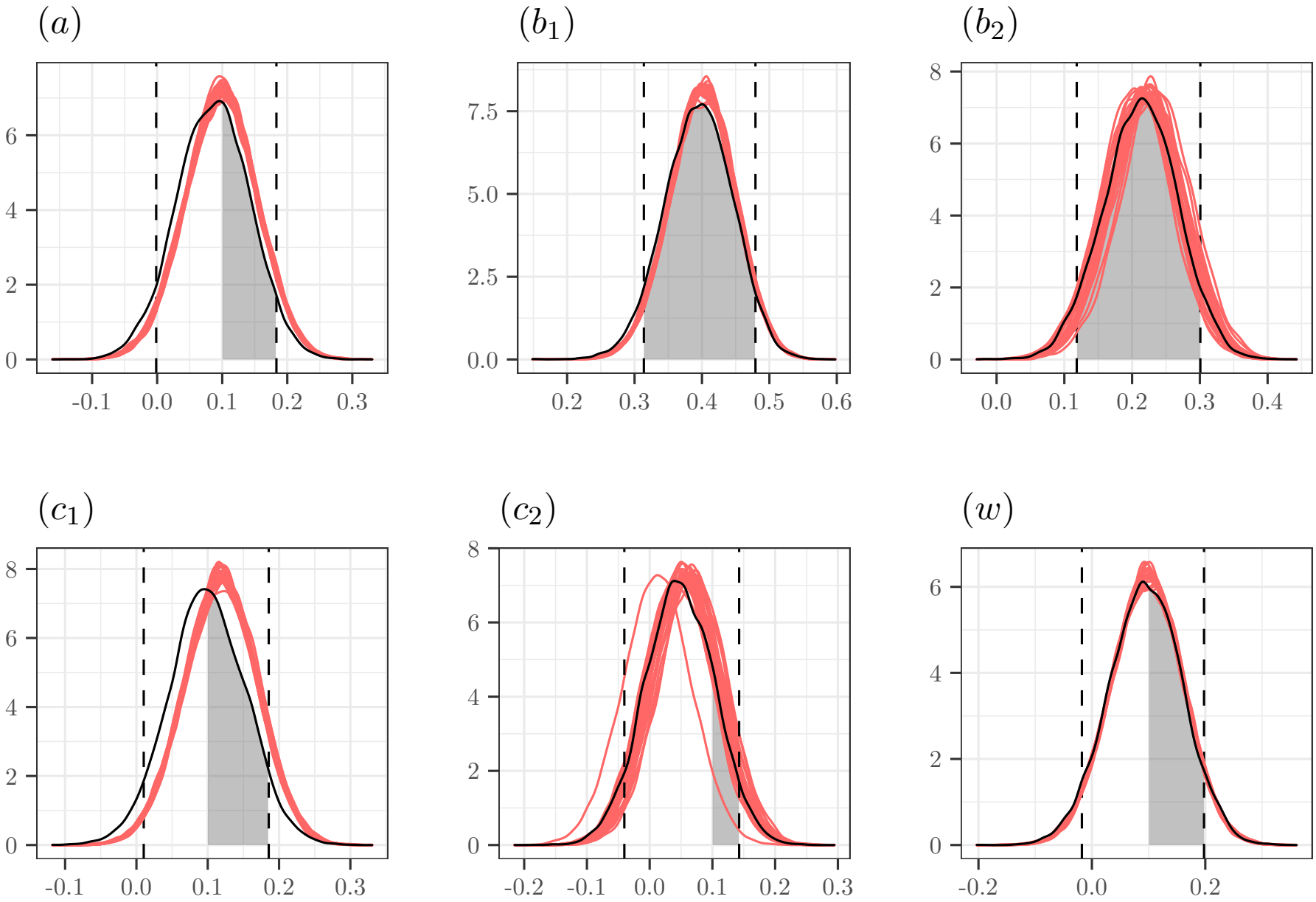
- Abbiamo inoltre definito l'intervallo $[-0.1, 0.1]$ come *Region of Practical Equivalence* (Kruschke, 2018).

Kruschke, J.K. (2018). Rejecting or accepting parameter values in Bayesian estimation. *Advances in Methods and Practices in Psychological Science*, 2, 270–280.

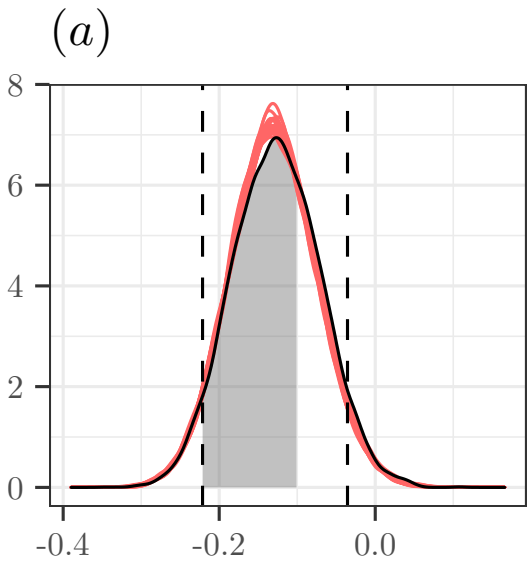
Permissive parenting

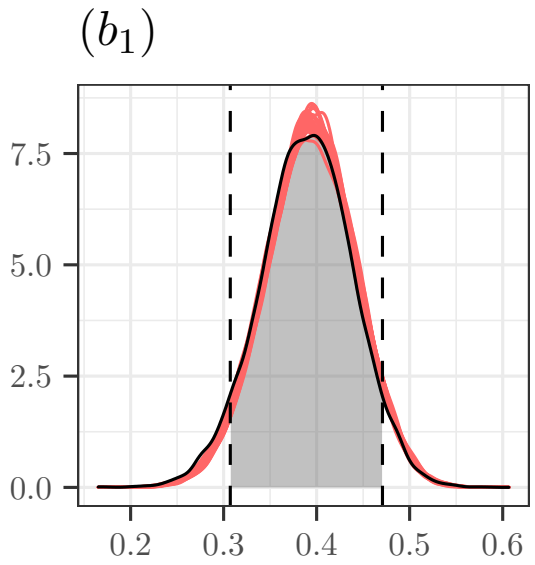


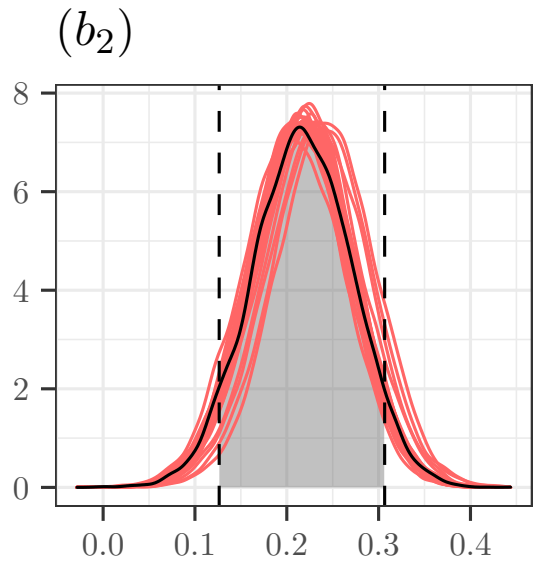
Authoritarian parenting

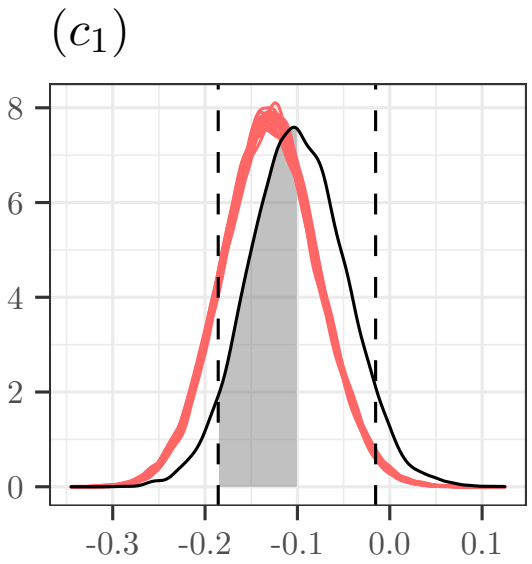


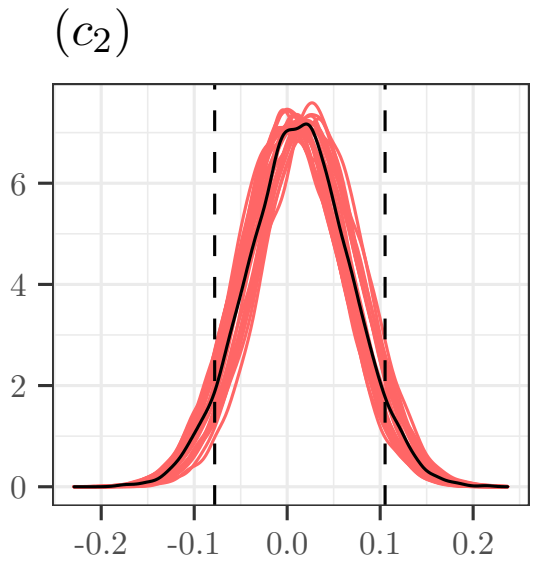
Authoritative parenting











Conclusioni

In generale:

- **Il problema dei *missing data* non è banale.**

In generale:

- **Il problema dei *missing data* non è banale.**

Nello specifico:

- Osservando i risultati appare che l'imputazione non produca grossi cambiamenti nelle stime.
- Comunque i pur minimi cambiamenti portano sempre ad "aumentare" le stime dei parametri.

In generale:

- **Il problema dei *missing data* non è banale.**

Nello specifico:

- Osservando i risultati appare che l'imputazione non produca grossi cambiamenti nelle stime.
- Comunque i pur minimi cambiamenti portano sempre ad "aumentare" le stime dei parametri.

Pertanto:

- Se l'imputazione non comporta dei cambiamenti, a che serve?

In generale:

- **Il problema dei *missing data* non è banale.**

Nello specifico:

- Osservando i risultati appare che l'imputazione non produca grossi cambiamenti nelle stime.
- Comunque i pur minimi cambiamenti portano sempre ad "aumentare" le stime dei parametri.

Pertanto:

- Se l'imputazione non comporta dei cambiamenti, a che serve?
- E se implica dei cambiamenti, come li dobbiamo interpretare?

massimiliano.pastore@unipd.it
http://dpss.psy.unipd.it/psicostat/
http://lilia.dpss.psy.unipd.it/~massimiliano.pastore/