Dati multi-fonte e analisi territoriali

Marco Tosi, Irene Barbiera, e Federico Gianoli

Dipartimento di Scienze Statistiche

Introduzione a «Risposte mancanti»

- **Risposte mancanti per l'intera unità/caso** (*unit non-response*) dovute alla non risposta di un caso o una unità statistica / mancato contatto con l'intervistato.
  - **In SHARE:** abbiamo informazione su coloro che sono stati campionati ma non hanno fatto l'intervista. Questo può avere conseguenze sulla rappresentatività del campione e quindi sull'inferenza.
- **Risposte mancanti in un singolo item** (*item non-response*) dovute alla non risposta in una singola variabile o item.
  - **In SHARE:** abbiamo visto in qualche caso che le variabili relative alla salute contengono dei valori mancanti.

Alcune definizioni

- Prendiamo Y una matrice (4 casi X 4 variabili) in cui supponiamo che i valori siano osservati (ossia la matrice dei valori veri). ***Y12 Y23 Y31*** e ***Y44*** sono i valori che in realtà sono mancanti.
- Prendiamo M come la matrice dei valori mancanti in cui abbiamo un indicatore uguale a 1 per le non risposte.

$$Y = \begin{bmatrix} y_{11} & \boldsymbol{y_{12}} & y_{13} & y_{14} \\ y_{21} & y_{22} & \boldsymbol{y_{23}} & y_{24} \\ \boldsymbol{y_{31}} & y_{32} & y_{33} & y_{34} \\ y_{41} & y_{42} & y_{43} & \boldsymbol{y_{44}} \end{bmatrix}$$

$$M = \begin{bmatrix} 0 & \boldsymbol{1} & 0 & 0 \\ 0 & 0 & \boldsymbol{1} & 0 \\ \boldsymbol{1} & 0 & 0 & 0 \\ 0 & 0 & 0 & \boldsymbol{1} \end{bmatrix}$$

Meccanismi sottostanti: MCAR

- **MCAR =** ***missing completely at random,*** ossia il meccanismo generativo non dipende da variabili osservate o non osservate. La distribuzione condizionata dei valori nella matrice M dati i valori di Y è $P(\boldsymbol{M}|\boldsymbol{Y}) = P(\boldsymbol{M})$. Ogni analisi che esclude tali valori mancanti rimane consistente benché poco efficiente.
  - **Esempio:** in uno studio longitudinale sui test fisici alcuni rispondenti non si prestano al test perché tra una wave e l'altra cambiano area di residenza. MCAR se la decisione di cambiare residenza è esogena alle informazioni rilevate nello studio.
- **MCAR** è solitamente una assunzione irrealistica e i dati osservati sono come un sotto-campione random del campione originario.

Meccanismi sottostanti: MAR

- **MAR = *missing at random***, ossia il meccanismo generativo può dipendere da qualche variabile osservata $P(\boldsymbol{M}|\boldsymbol{Y}) = P(\boldsymbol{M}|\boldsymbol{Y}_{\boldsymbol{obs}})$. La probabilità di avere valori mancanti in Y può dipendere da X ma non da Y stesso o da meccanismi non osservabili.
  - ***Esempio:*** la probabilità di non riportare il proprio reddito può dipendere dallo stato civile (es., *partnership premium / penalty*) ma non dal fatto di avere un reddito alto o basso.
  - Intervistati che lasciano lo studio perché hanno avuto effetti collaterali del trattamento (X) dello studio e quindi non si sottopongono alla misurazione del test (Y).

  - MAR è solitamente l'assunzione che facciamo quando imputiamo i valori mancanti attraverso tecniche di imputazione multipla.

Meccanismi sottostanti: NMAR

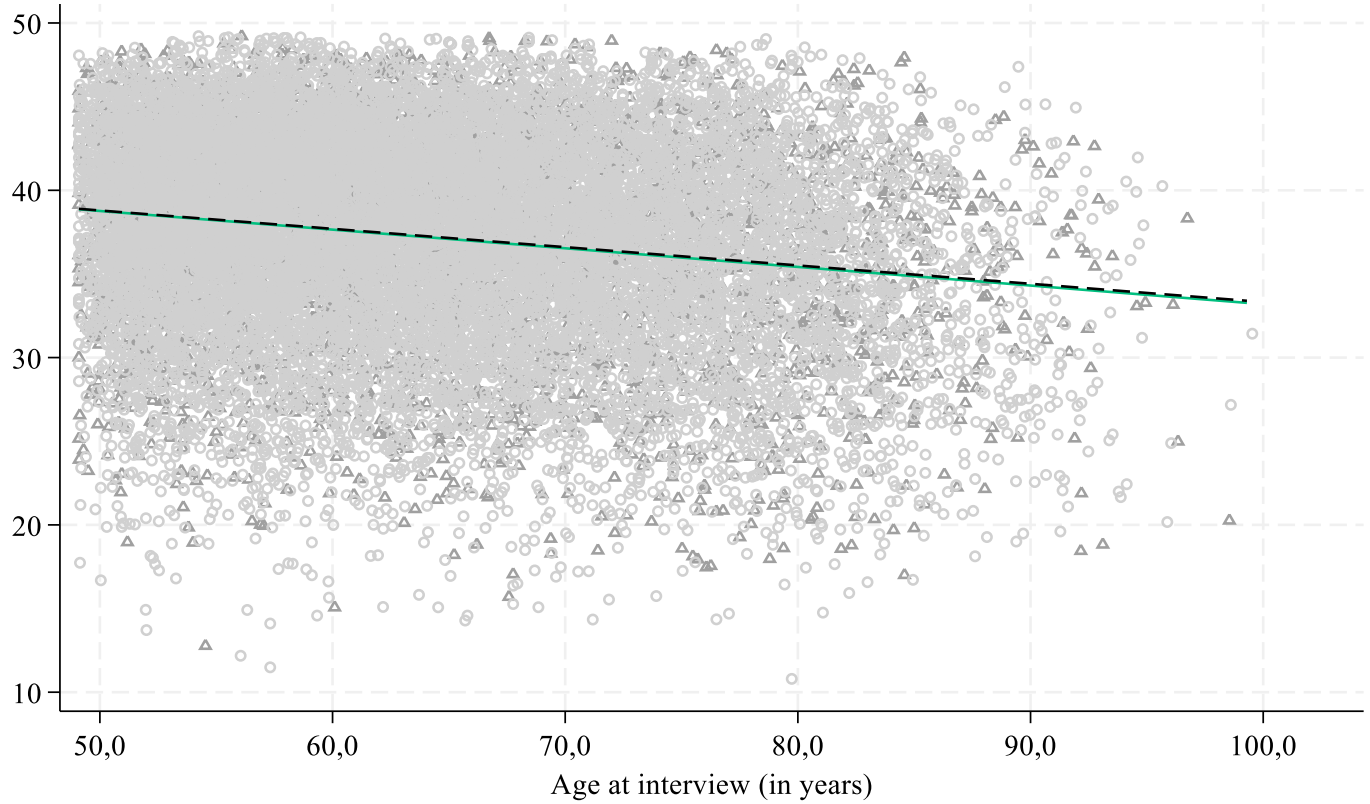
- **NMAR = Not missing at random**, ossia il meccanismo generativo può dipendere da variabili osservate e non osservate $P(\boldsymbol{M}|\boldsymbol{Y}) \neq P(\boldsymbol{M}|\boldsymbol{Y}_{\boldsymbol{obs}})$. La probabilità di avere valori mancanti in Y può dipendere da Y stesso.
  - **Esempio:** I più ricchi hanno più probabilità di non riportare il proprio reddito, per questioni legate all'evasione fiscale, oppure i più poveri potrebbero essere più inclini a non farsi vedere come tali durante una intervista.
  - Ragioni etiche possono essere il meccanismo sottostante a NMAR: come non sottoporsi al test della pressione del sangue (Y) perché si hanno valori molto alti in Y.

Missing MCAR, MAR, e MNAR

Figura 1 – Rappresentazione grafica di (a) missing completely at random (MCAR), (b) missing at random (MAR) e (c) missing not at random (MNAR) (Schafer e Graham, 2002)

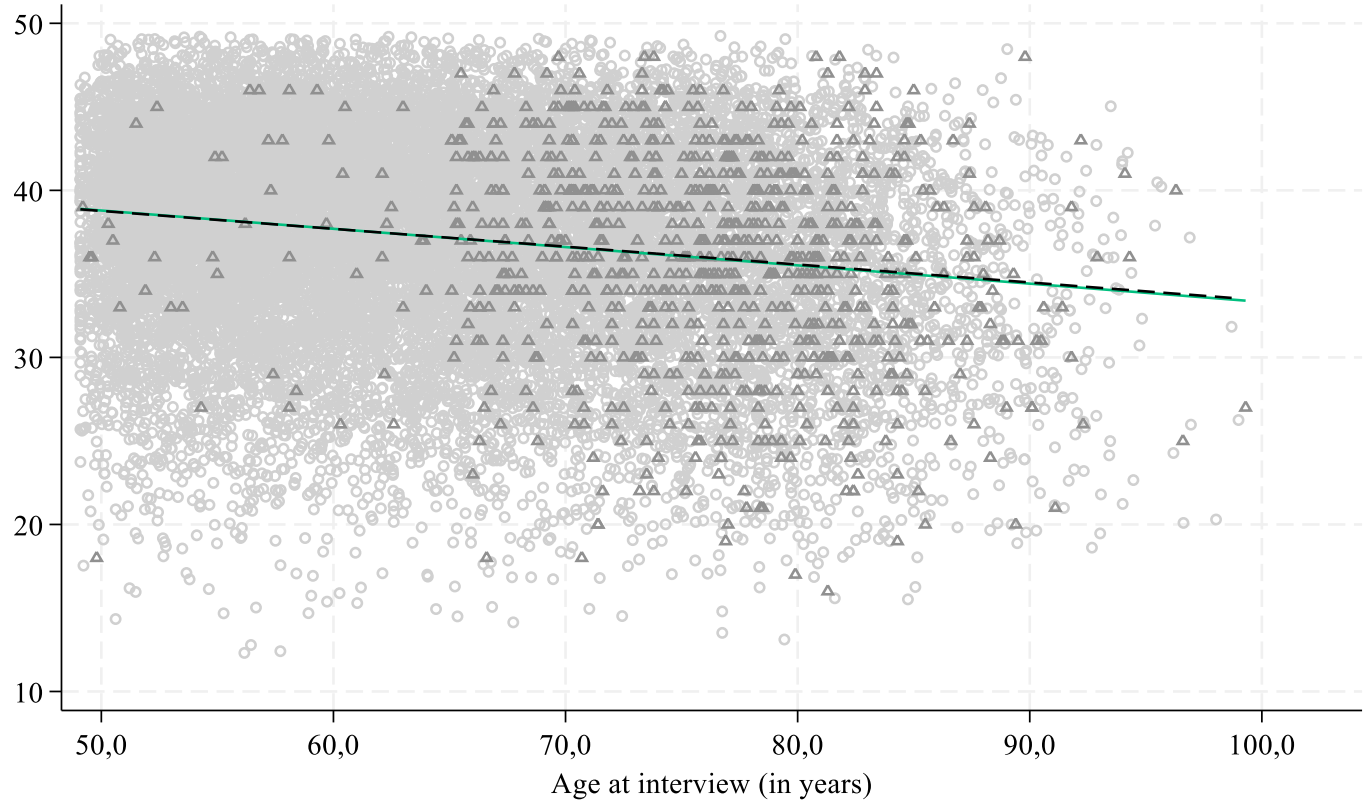
MCAR

- Y=Casp qualità della vita (25% di dati mancati casualmente distribuiti). Retta verde escludendo i dati mancanti.



MAR

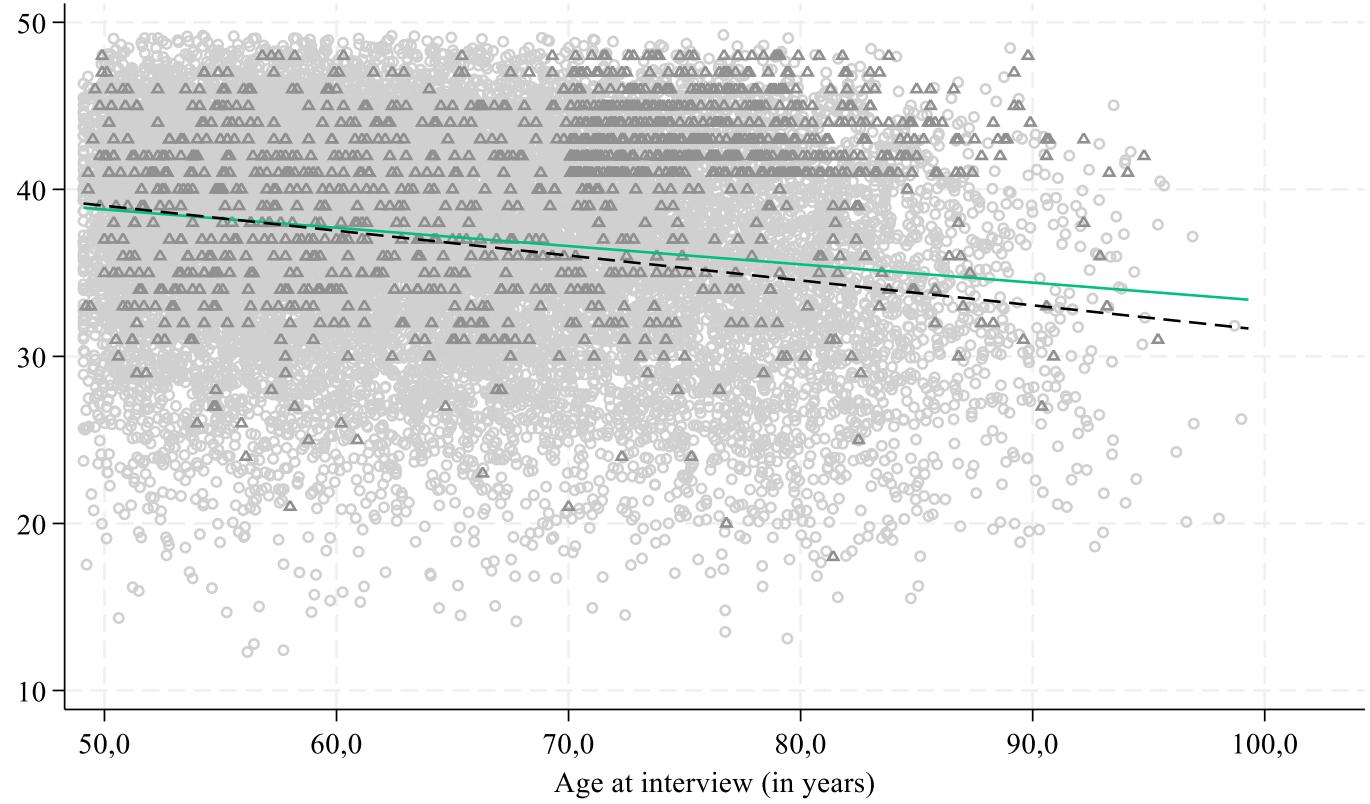
- Y=Casp qualità della vita (25% di dati mancati scelti casualmente all'interno di gruppi di età).



Possiamo ricostruire la distribuzione condizionata di Y partendo dai valori osservati di Y per ogni livello di X.

NMAR

- Y=Casp qualità della vita (casi mancanti correlati a Y e X). Claster di dati mancanti per età avanzate e alti valori di Casp



Tipi di missing:
1- Risposte mancanti per costruzione

- *1.1 Inclusione di filtri:* solo alcuni intervistati rispondono a quella domanda
  - *Esempio:* informazioni sui redditi da lavoro solo per coloro che hanno una posizione occupazionale
- *1.2* Nei dati longitudinali/panel alcune informazioni vengono chieste solo durante la prima intervista
  - *Esempio:* Informazioni costanti nel tempo, es, sesso, non chieste nella wave 2
- *1.3* Nei dati con struttura gerarchica alcune informazioni vengono chieste solo ad un rispondente per tutti i membri di una unità
  - *Esempio:* il capo famiglia risponde sui redditi del nucleo famigliare

Nei dati SHARE…

- Risposte mancanti dovute ai cambiamenti/ non cambiamenti nel tempo: durante la prima intervista si rileva lo stato civile, nelle successive solo i cambiamenti

	mergeid	wave	dn014_	dn044_
1	AT-000327-01	1	Married and living together with spouse	.
2	AT-000327-01	2	.	No, marital status has not changed
3	AT-000327-02	1	Married and living together with spouse	.
4	AT-000327-02	2	.	No, marital status has not changed
5	AT-001816-01	1	Married and living together with spouse	.
6	AT-001816-01	2	.	.
7	AT-001816-02	1	Married and living together with spouse	.
8	AT-001816-02	2	.	No, marital status has not changed

Nei dati SHARE…

- Un rispondente («capofamiglia») riporta informazioni che riguardano tutti i membri del nucleo (vivere in una casa di proprietà/ in affitto)

	mergeid	hhid1	hou_resp	ho002_
55296	AT-000327-01	AT-000327-A	Not household respondent	.
55297	AT-000327-02	AT-000327-A	Household respondent	Tenant
55298	AT-001816-02	AT-001816-A	Not household respondent	.
55299	AT-001816-01	AT-001816-A	Household respondent	Owner
55300	AT-002132-03	AT-002132-A	Not applicable	.
55301	AT-002132-08	AT-002132-A	Not applicable	.
55302	AT-002132-01	AT-002132-A	Household respondent	Subtenant
55303	AT-002132-07	AT-002132-A	Not applicable	.
55304	AT-002132-02	AT-002132-A	Not applicable	.
55305	AT-002132-06	AT-002132-A	Not applicable	.

Missing per costruzione dei dati longitudinali

- Riportare informazioni da una wave a quella successiva

	mergeid	wave	dn014_	marital_stat	dn044_
1	AT-000327-01	1	Married and living together with spouse	married	.
2	AT-000327-01	2	.	married	No, marital status has not changed
3	AT-000327-02	1	Married and living together with spouse	married	.
4	AT-000327-02	2	.	married	No, marital status has not changed
5	AT-001816-01	1	Married and living together with spouse	married	.
6	AT-001816-01	2	.	married	.
7	AT-001816-02	1	Married and living together with spouse	married	.
8	AT-001816-02	2	.	married	No, marital status has not changed

Missing nei dati longitudinali

- Problemi nella costruzione dei dati longitudinali: rispondenti che non vengono intervistati nella prima wave, entrano nella seconda, ma rispondono alle domande sui cambiamenti/ non cambiamenti nel tempo

	mergeid	wave	interview	marital_stat	dn044_
95	AT-017298-02	1	No interview	.	.
96	AT-017298-02	2	Main interview	.	No, marital status has not changed
261	AT-057665-01	1	Main interview	.	.
846	AT-220315-02	2	Main interview	.	No, marital status has not changed
982	AT-262986-01	2	Main interview	.	No, marital status has not changed
1232	AT-322004-01	2	Main interview	.	No, marital status has not changed
1376	AT-366334-02	1	No interview	.	.
1377	AT-366334-02	2	Main interview	.	No, marital status has not changed
2086	AT-543275-01	2	Main interview	.	No, marital status has not changed
2205	AT-575677-01	2	Main interview	.	No, marital status has not changed
2680	AT-707856-01	1	Main interview	.	.

Una soluzione… parziale

- Possiamo imputare le informazioni che riporta il partner attraverso l'ID del partner e l'ID della famiglia. Se 2 rispondenti vivono nello stesso nucleo ma solo uno dei 2 riporta lo stato civile

	mergeid	IDfam	wave	interview	marital_stat	mergeidp1	mergeidp2
93	AT-017298-02	AT-017298-A	1	No interview	.	AT-017298-01	
94	AT-017298-01	AT-017298-A	1	Main interview	married	AT-017298-02	
95	AT-017298-01	AT-017298-A	2	Main interview	married		AT-017298-02
96	AT-017298-06	AT-017298-A	2	No interview	.		
97	AT-017298-02	AT-017298-A	2	Main interview	.		AT-017298-01
98	AT-017298-07	AT-017298-A	2	No interview	.		

Stato civile del partner

- Quando il partner riporta lo stato civile «coniugato» o «coppia di fatto» possiamo ragionevolmente imputare la variabile
- Ma per i divorziati e i vedovi si può trattare di una nuova relazione di coppia e non sappiamo se sono entrambi vedovi o divorziati.
- Imputando lo stato civile delle coppie abbiamo lo 0.58% di missing

	mergeid	IDfam	wave	interview	marital_stat	IDpartner	partn
45535	FR-452543-01	FR-452543-A	2	Main interview	married	FR-452543-02	.
45773	FR-480740-02	FR-480740-A	2	Main interview	.	FR-480740-01	6
45774	FR-480740-01	FR-480740-A	2	Main interview	widowed	FR-480740-02	.
45788	FR-482124-01	FR-482124-A	1	Main interview	divorced	FR-482124-02	.
45789	FR-482124-02	FR-482124-A	1	Main interview	.	FR-482124-01	5
45790	FR-482124-01	FR-482124-A	2	Main interview	divorced	FR-482124-02	.
45791	FR-482124-02	FR-482124-A	2	Main interview	.	FR-482124-01	5

Chi sono coloro che non rispondono?

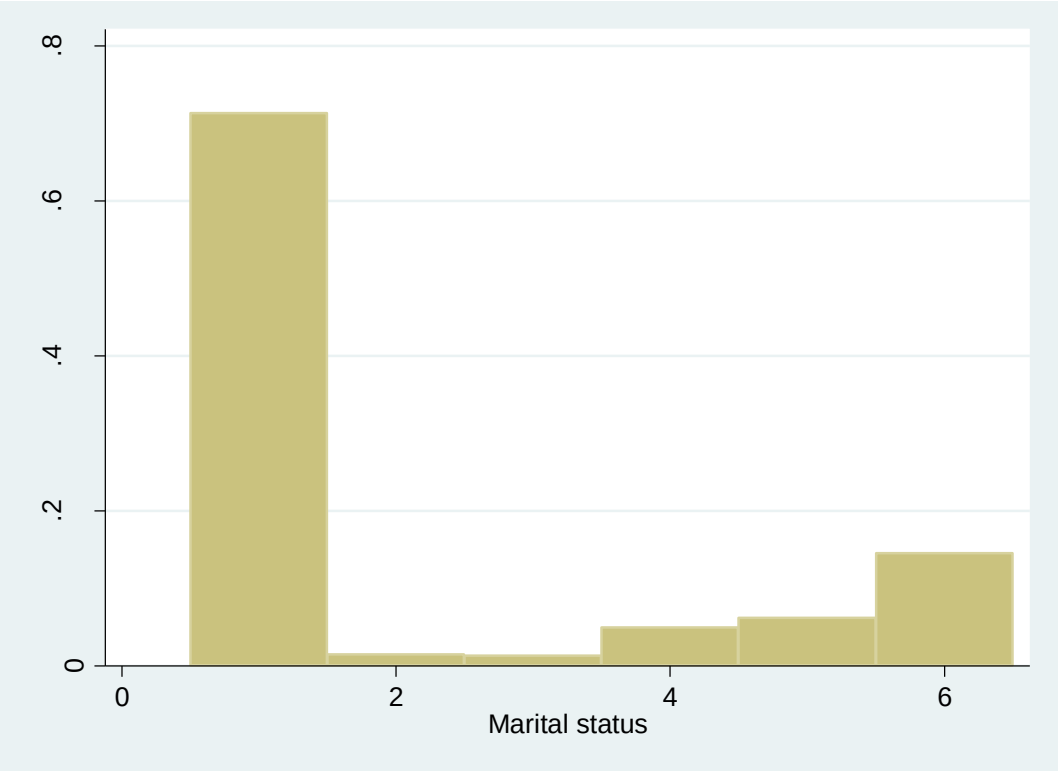
- Hanno mediamente un’età di 66 anni contro i 64 dei rispondenti
- Sono sovra-rappresentati in Israele (anche 82 veterani ebrei). In una analisi si potrebbe pensare di escludere Israele in quanto unico stato extra europeo
- Netta maggioranza nella wave 2 rispetto che nella wave 1.
- Hanno (leggermente) più limitazioni dovute allo stato di salute.

- *Notare che* recuperiamo dei casi ma la distribuzione cambia solo in minima parte rispetto alla variabile originaria

Distribuzione

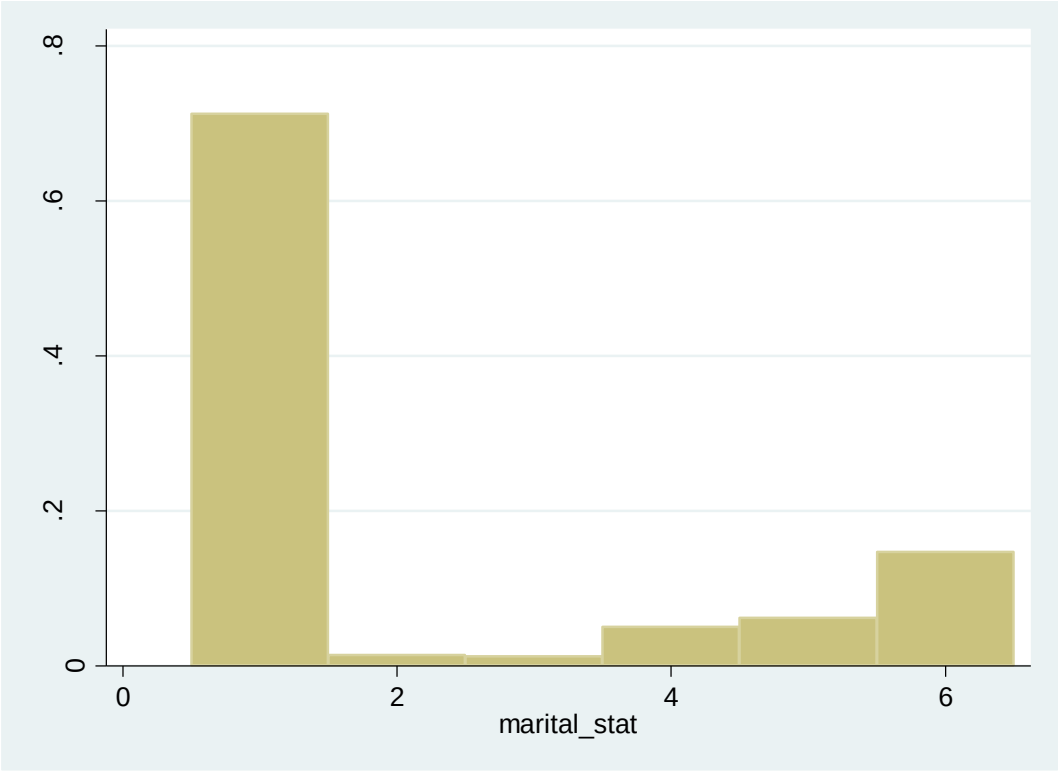
Variabile osservata

Esclusione di tutti i missing



Variabile imputata

Esclusione dei missing «veri»



2- Risposte mancanti «vere»

- Abbiamo Molti metodi per affrontare il problema dei dati mancanti. In questo corso verranno raggruppati in Approcci «Tradizionali» e Approcci «Moderni».
- Il metodo più appropriato dipende dal **meccanismo generativo dei missing** (es., assunzioni: MAR) e **dalla distribuzione e dal pattern** dei valori mancanti.

Il pattern delle risposte mancanti

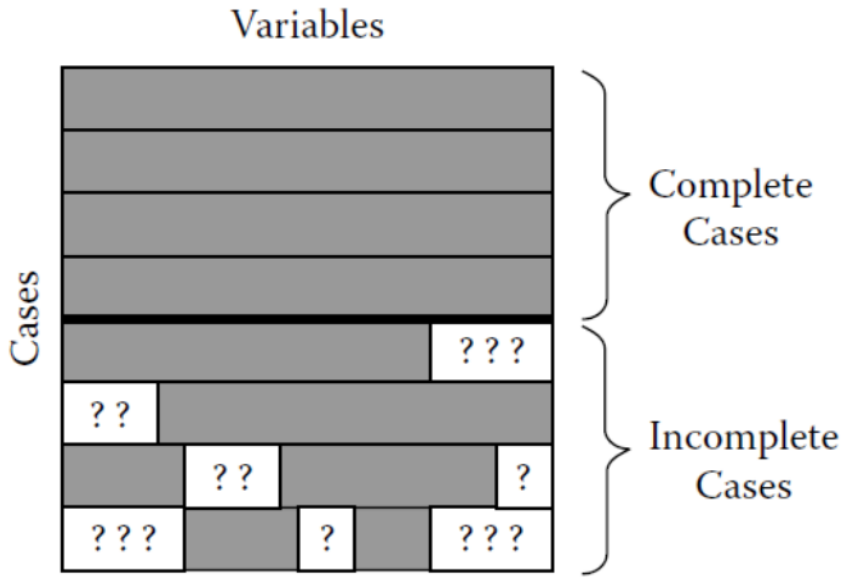
- Prendiamo Y come se fosse una matrice in cui tutti i valori sono osservati. In grassetto quei valori che in realtà sono mancanti.
- Abbiamo quindi anche la matrice M in cui 1 indentifica i valori mancanti e 0 quelli osservati nella matrice Y
- I pattern delle risposte mancanti sono le distribuzioni osservate in M.

$$Y = \begin{bmatrix} y_{11} & \mathbf{y_{12}} & y_{13} & y_{14} \\ y_{21} & y_{22} & \mathbf{y_{23}} & y_{24} \\ \mathbf{y_{31}} & y_{32} & y_{33} & y_{34} \\ y_{41} & y_{42} & y_{43} & \mathbf{y_{44}} \end{bmatrix}$$

$$M = \begin{bmatrix} 0 & \mathbf{1} & 0 & 0 \\ 0 & 0 & \mathbf{1} & 0 \\ \mathbf{1} & 0 & 0 & 0 \\ 0 & 0 & 0 & \mathbf{1} \end{bmatrix}$$

Il pattern delle risposte mancanti

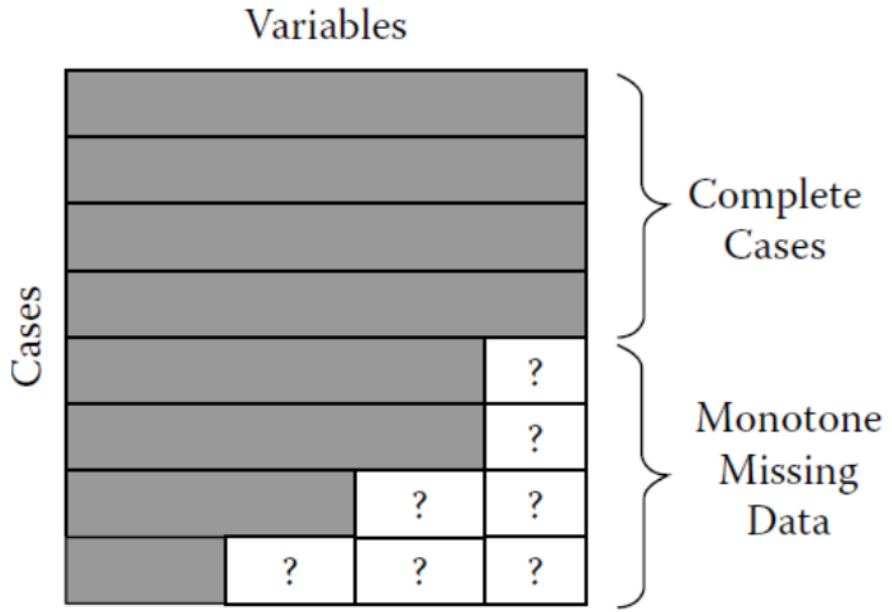
- Quando il pattern è **non-monotono**, M si presenta così:



- Questo pattern può essere dovuto a rifiuti a rispondere o «non so». Si ricorre spesso a tecniche di imputazione multipla.

Il pattern delle risposte mancanti

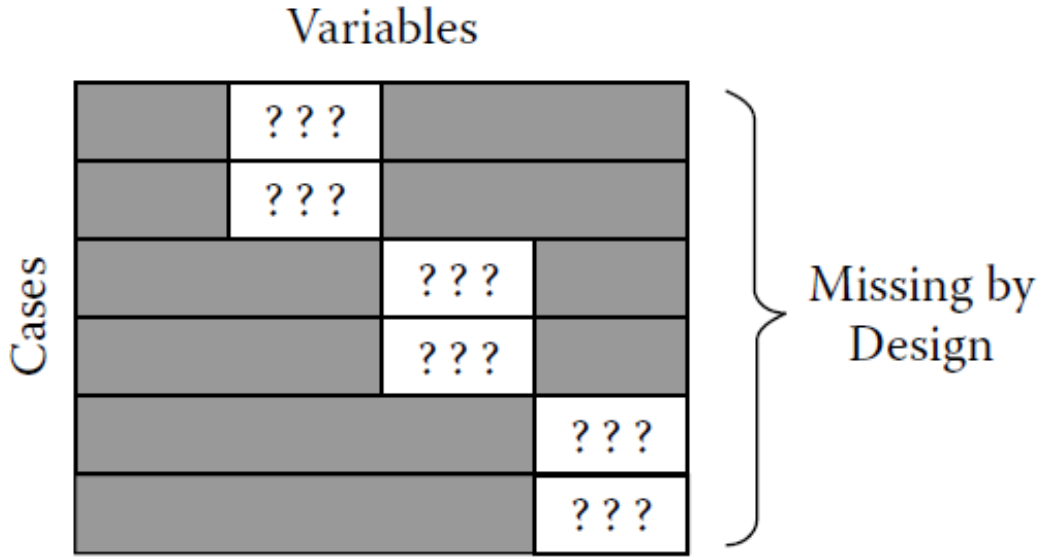
- Quando il pattern è **monotono**, M si presenta così:



- In questo caso i pesi di non risposta possono essere più efficaci dell'imputazione. Imputazione è più semplice perché X2 è missing quando X1 è missing (anche se può avere altri missing non correlati a X1)

Il pattern delle risposte mancanti

- Quando il pattern è **strutturato ma non-monotono**, M si presenta così:



- Questo pattern può essere dovuto all'assegnazione random di alcuni quesiti del questionario.

Perché dovremmo preoccuparci dei valori mancanti («veri»)?

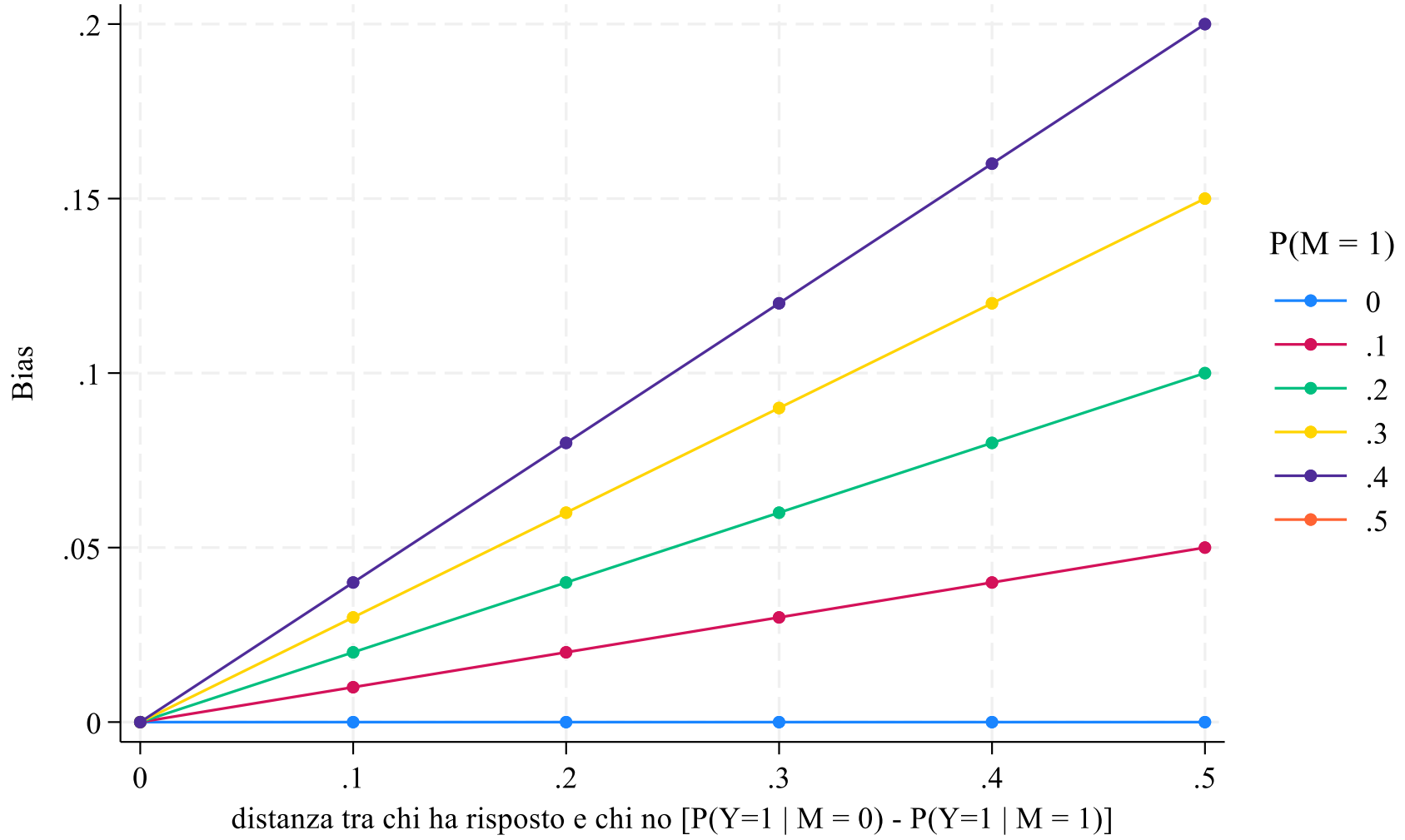
- Se assumiamo che il meccanismo generativo dei valori mancanti sia MAR o NMAR le nostre stime sono distorte.
- Se abbiamo una variabile dummy Y e siamo interessati a stimare le probabilità con cui si verifica Y=1 (e M=1 quando Y è missing), allora abbiamo (per il teorema della probabilità totale):

$$P(Y=1) = P(Y=1 \mid M=0) * P(M=0) + P(Y=1 \mid M=1) * P(M=1)$$

Ignorando i valori mancanti, la stima di P(Y=1) è P(Y=1 | M = 0), ottenendo un BIAS = $P(M = 1) * [P(Y=1 \mid M = 0) - P(Y=1 \mid M = 1)]$

Che quindi *dipende dalla proporzione dei missing e dalla differenza tra il valore di P(Y=1) tra coloro che hanno risposto e chi non ha risposto*.

Identificazione del bias



Identificazione del range del bias

Dato che: $0 \leq P(Y=1 \mid M = 1) \leq 1$

$LB = P(Y=1 \mid M=0) * P(M=0)$

$UB = P(Y=1 \mid M=0) * P(M=0) + P(M=1)$

P(M=1) come misura dell'incertezza attorno a P(Y=1) dovuta alle risposte mancanti. Quindi la probabilità con cui si verifica un dato mancante viene qui usata come misura di distorsione.

Risposte mancanti «vere»
Approcci tradizionali o deterministici

1- Esclusione dei casi con risposte mancanti

2- Sostituire con la media

2.1- Sostituire con la media di gruppo

3- Creare un indicatore per i missing

4- Campionamento aleatorio

1- Eliminare i casi con risposte mancanti

- Assunzione che il meccanismo generativo delle risposte mancanti non dipende dalle variabili osservate o non osservate (**MCAR**). Spesso viene violata nella pratica.
- Anche quando la distribuzione dei dati mancanti è (verosimilmente) casuale (MCAR), la riduzione della numerosità è sufficiente a provocare una **perdita di efficienza. + Errore di TIPO II**
- Se l'assunzione MCAR non è supportata le stime possono essere distorte in entrambe le direzioni, in quanto il campione non è più rappresentativo della popolazione (problemi di validità esterna).

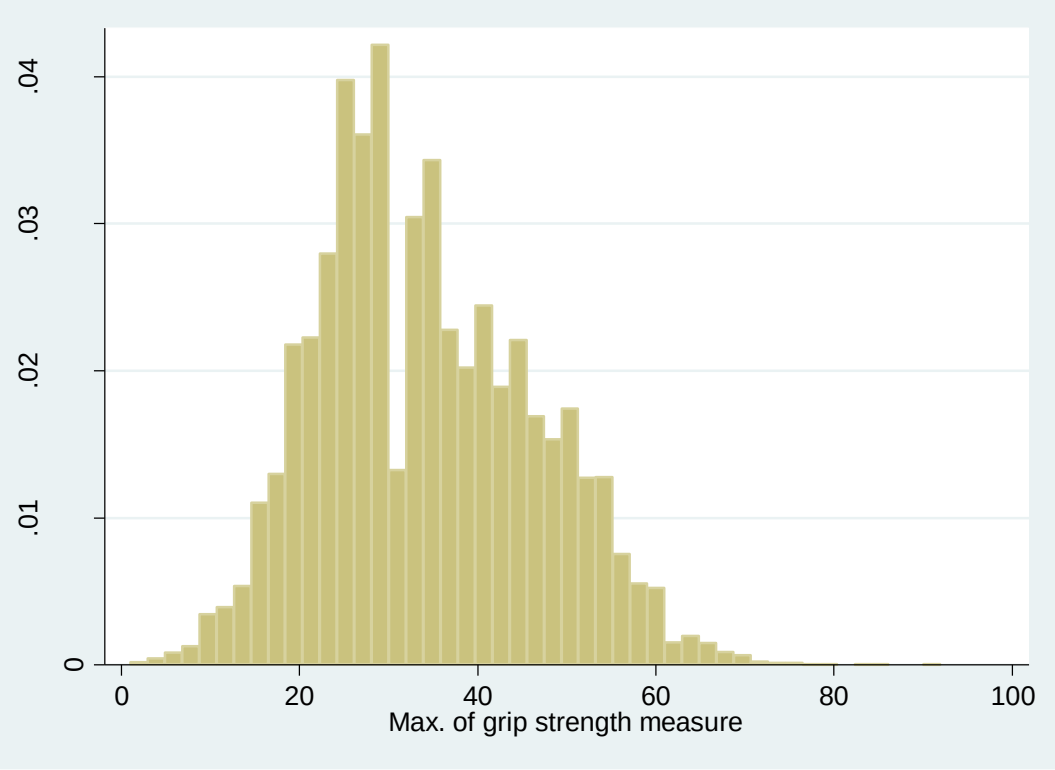
2- Sostituzione con la media

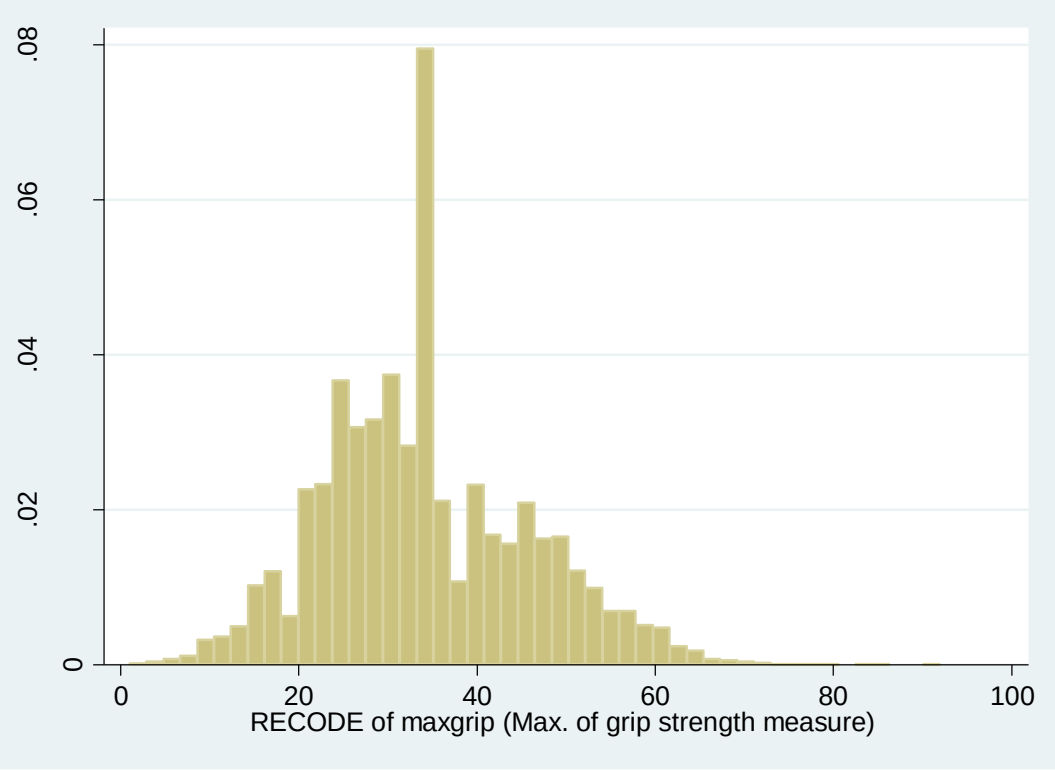
- *Assunzioni:* risposte mancanti sono osservazioni casuali di una distribuzione normale. MISSING AT RANDOM
- Tuttavia gli estremi potrebbero avere propensione maggiore a non rispondere
  - ESEMPIO: Fasce basse e alte di reddito sono più reticenti a rispondere a domande sui loro redditi.
- *Distribuzione distorta:* riduzione (artificiosa) della variabilità, più concentrata attorno alla media, minore varianza.
  - *Sd MAXGRIP: 12.11*
  - *Sd GRP2 (no missing): 11.55*

Forza della presa della mano (in SHARE)

- Variabile originale (6110 missing)

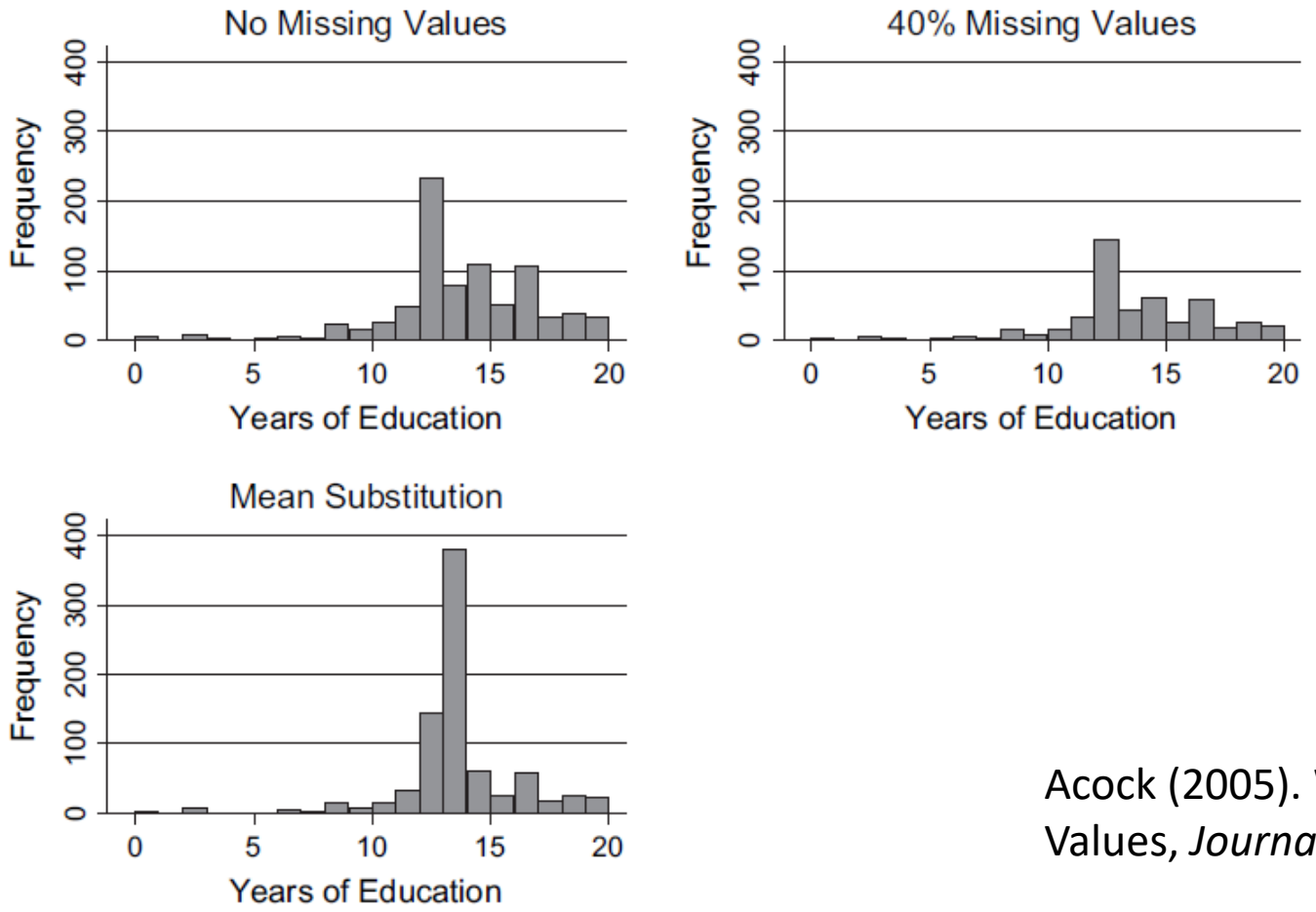
sostituzione con la media





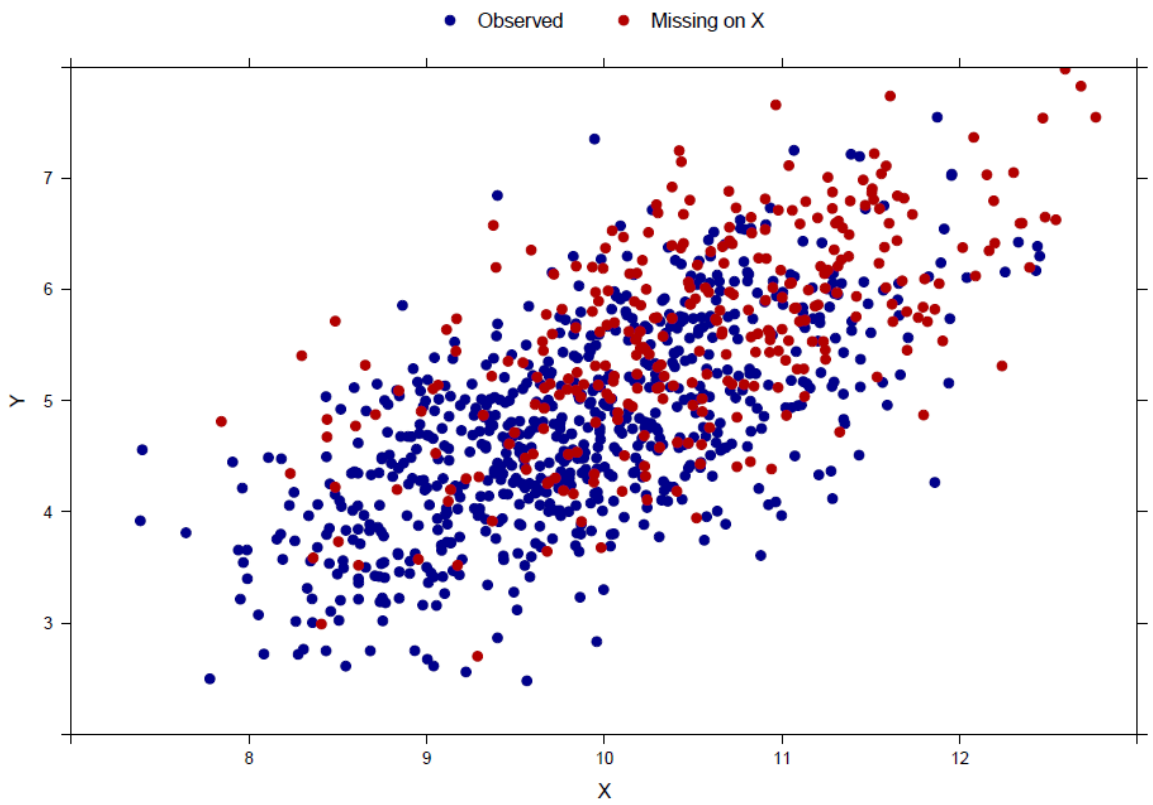
Sostituzione con la media (un esempio)

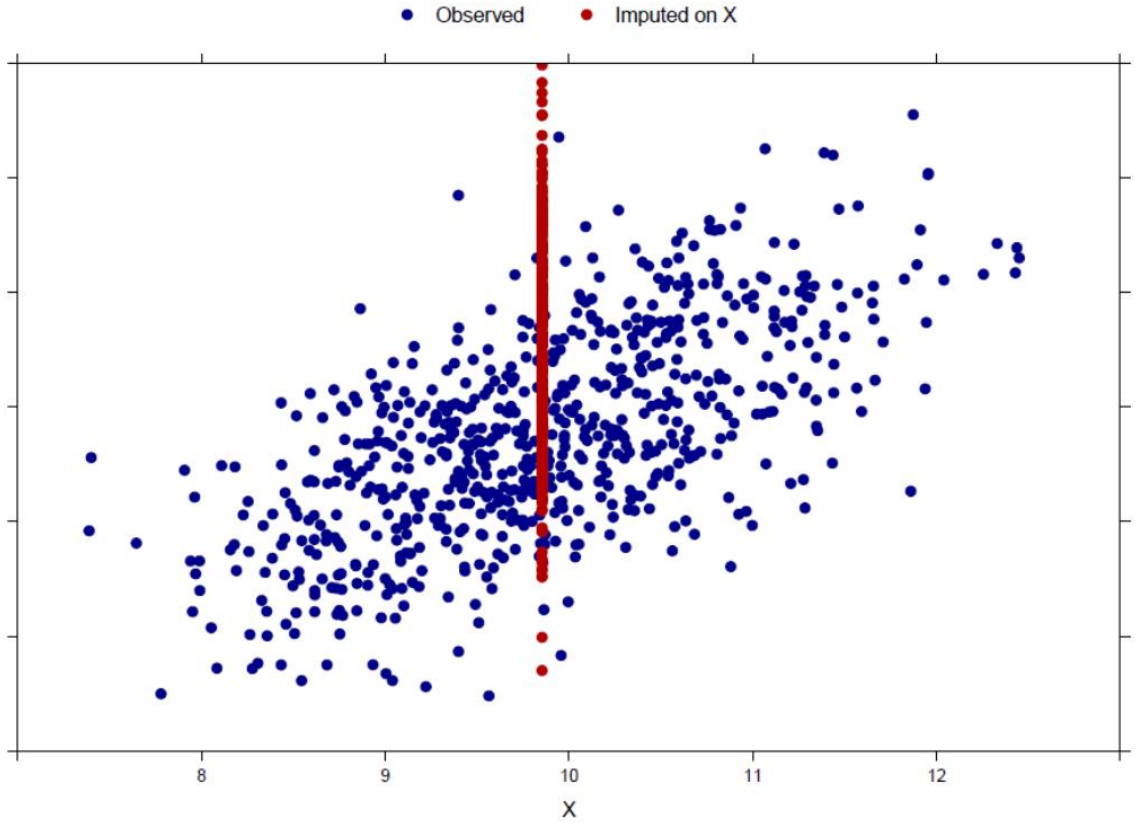
FIGURE 1. MEAN SUBSTITUTION DISTORTS DISTRIBUTION AND ATTENUATES VARIANCE



Acock (2005). Working With Missing Values, *Journal of Marriage and Family*

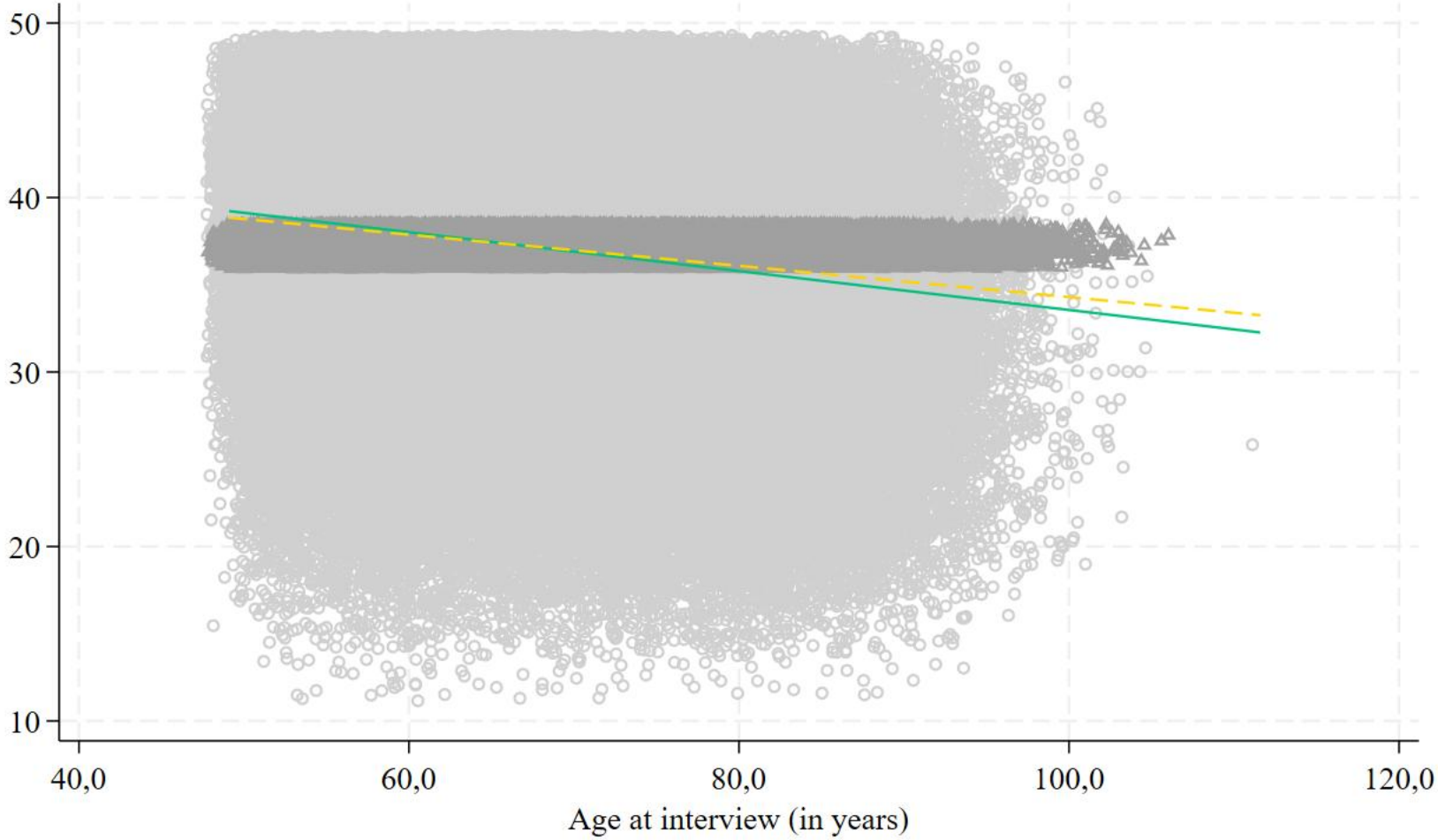
Sostituzione con la media (un esempio)





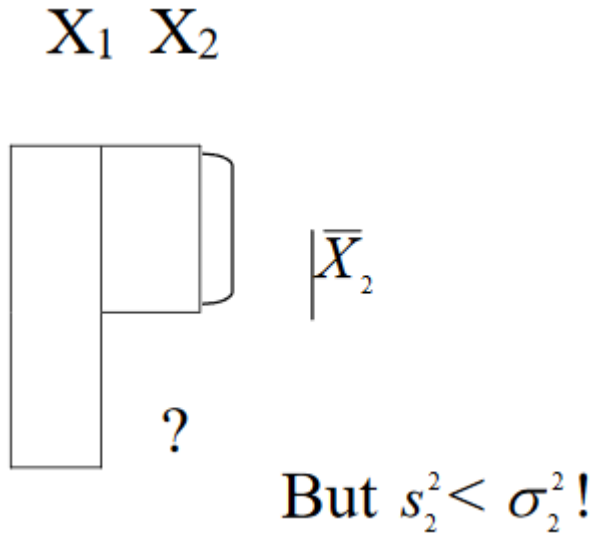
Sostituzione con la media (un esempio)

- Y= CASP; in giallo la retta che avremmo sostituendo con la media.



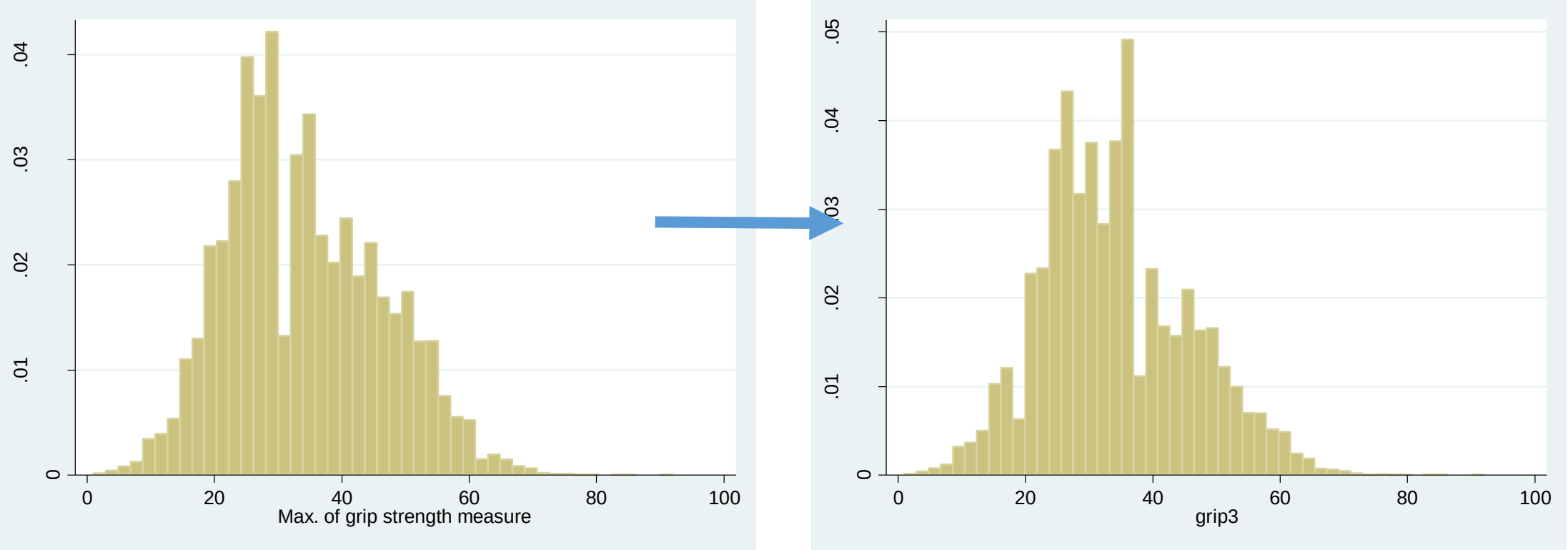
Sostituzione con la media: limiti

- Introduce una distorsione nella distribuzione della variabile con un picco artificiale sulla media.
- Non dà buoni risultati per la stima della varianza
- Provoca distorsioni nella relazione tra variabili
- Ricordate però che il BIAS dipende anche da P(M=1)



2.1 Sostituzione con la media di gruppo

- Stime migliori per i gruppi definiti dalla variabile
- Varianza maggiore (più simile a quella reale/ originale)

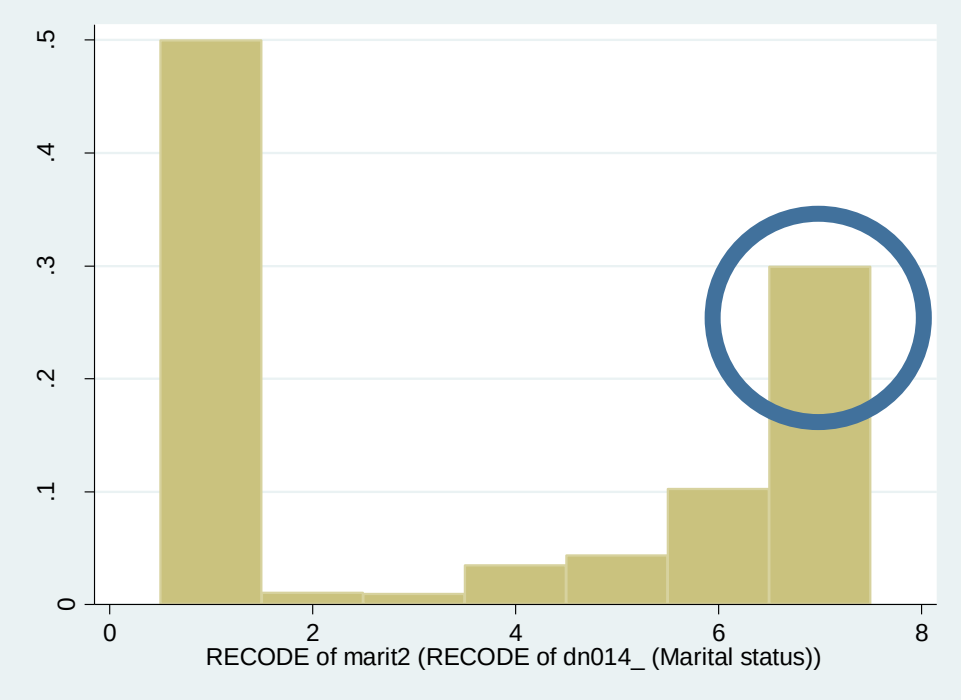


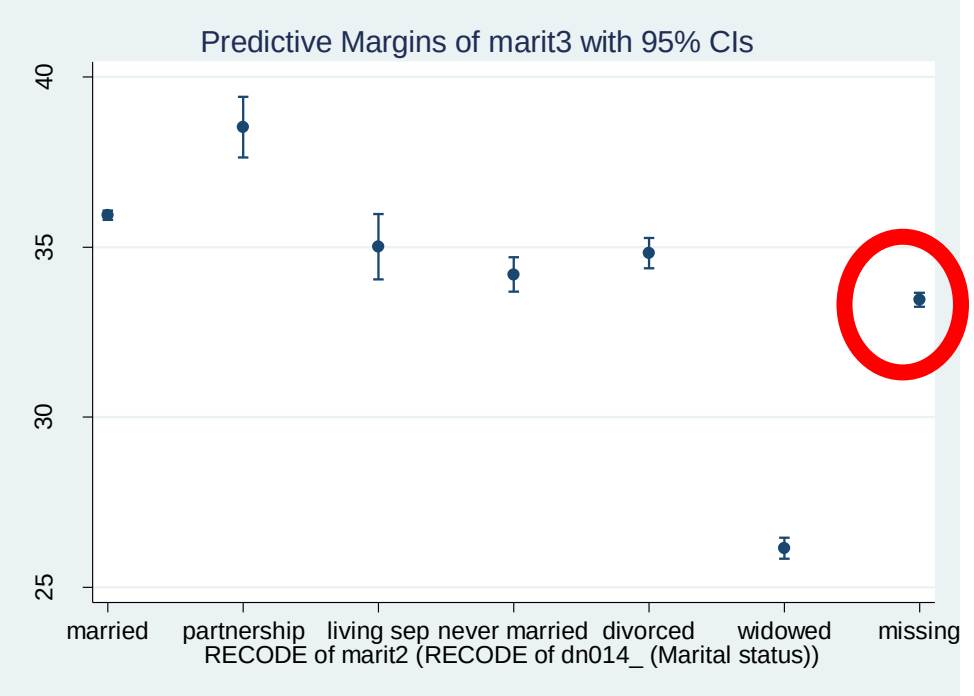
Sostituzione con la media di gruppo: limiti

- Introduce distorsioni (sebbene in maniera meno evidente del metodo precedente) nella distribuzione della variabile, creando una serie di picchi artificiali in corrispondenza della media di ciascuna classe.
- Provoca un’attenuazione della varianza della distribuzione dovuta al fatto che i valori imputati riflettono solo la parte di variabilità tra le classi (between) ma non quella all’interno delle classi (within).
- Provoca distorsioni nelle relazioni tra le variabili non considerate per la definizione delle classi di imputazione.

3- creare un indicatore per i missing «veri»

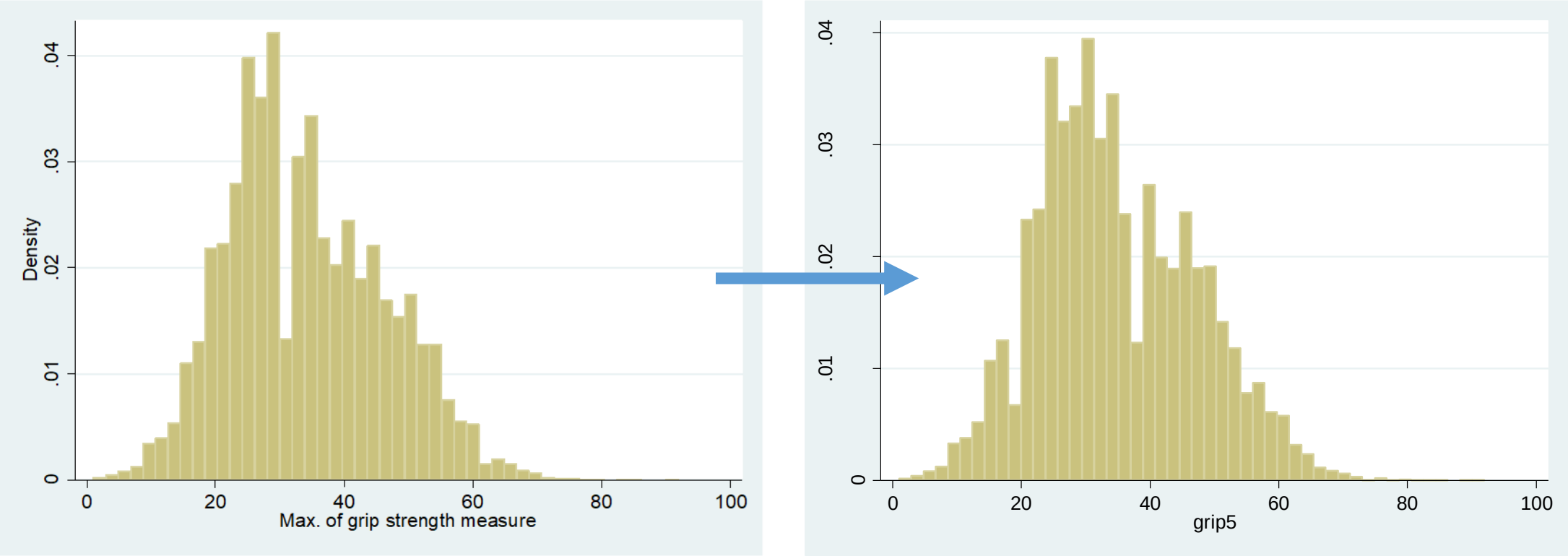
- Includiamo nelle analisi una categoria per i valori mancanti.
  - Problema di collinearità visto che coloro che non rispondono ad una domanda tendono a non rispondere anche ad altri items.
  - Distorsione delle stime quando includiamo più variabili indipendenti





4- campionamento aleatorio

- *Campionamento aleatorio:* valori mancanti imputati come se fossero distribuiti normalmente



Chi sono?
Ossia testare i meccanismo generativo

- Coloro che non fanno il test sulla forza di presa della mano sono tendenzialmente (rispetto a coloro che fanno il test):
  - Donne vedove con più limitazioni dovute alla salute (20% tra gli Irlandesi)
  - Missing completely at random è una assunzione che spesso non è supportata dai dati
  - Possiamo assumere MAR? (altre variabili sulle condizioni di salute ci fanno credere che siano distribuiti in modo MNAR).

Come cambiano il coefficiente dei vedovi

- OLS: forza della mano = B(stato civile) + B(wave) + B(gender) + constante

Sulla Indipendente: Sulla Dipendente:	No imputazione	Missing per costruzione	Missing indicator	Sostituzione Missing con moda
No imputazione	-4.996 (0.129)	-4.851 (0.107)	-4.949 (0.127)	-4.836 (0.107)
Mean substitution	-3.977 (0.119)	-3.767 (0.098)	-3.957 (0.117)	-3.759 (0.098)
Group mean substitution	*-5.357 (0.117)*	-5.212 (0.097)	-5.303 (0.116)	-5.196 (0.097)
Uniform random values	-3.213 (0.166)	*-2.793 (0.139)*	-3.202 (0.166)	-2.822 (0.140)
Normal random values	-3.395 (0.133)	-3.179 (0.110)	-3.387 (0.131)	-3.188 (0.110)