

Qualità - 3. Accordo.

Stefano Mizzaro

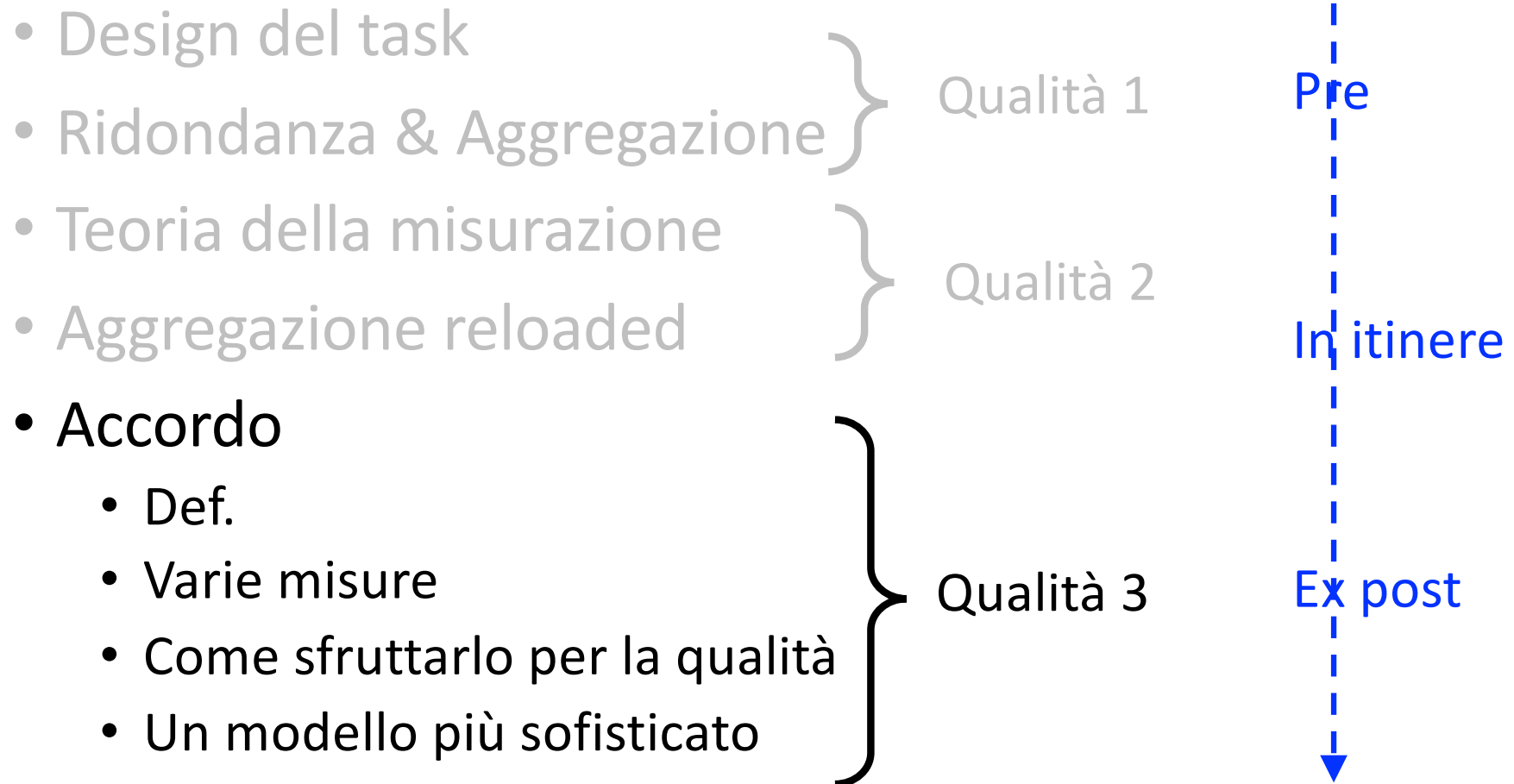
DMIF – UNIUD

mizzaro@uniud.it

Social Computing – Lezione 23

16 dicembre 2024

Schema



Riassunto Qualità 1 & 2: Tecniche per migliorare / ottenere buona qualità

- Metodologiche
 - Design del task (Pairwise, scale, presentazione, ...)
- In itinere
 - Test di qualificazione (Qualification test)
 - Controlli sintattici
 - Test nascosti (Gold questions, honey pots, coerenza)
 - Monitoraggio del comportamento (tempo, azioni, log, ...)
- Ridondanza. Aggregare
 - Il problema dell'aggregazione
 - Cenni di teoria della misura
- Ex-post
 - Pulizia dati (rimozione outlier, ecc.)
 - Aggregazione sofisticata (co-determination algorithms)
 - Sfruttare l'accordo

Obiettivi della lezione di oggi

- Continuare a discutere della qualità (parte 3)
- Ex post
- Accordo (Agreement)
 - Def. intuitive
- Sfruttare l'accordo come proxy per la qualità
 - Per l'affidabilità dei dati raccolti
 - Varie misure di accordo
 - Per la qualità del singolo worker
 - Un modello più sofisticato

Controlli ex post (già visto)

- Anche se il worker ha finito il lavoro e sottomesso i risultati, posso decidere di:
 - Dare pesi diversi ($<$, $>$) ai risultati di quel worker rispetto agli altri
 - Non usarli
 - Non accettarli e/o non pagarlo
 - Oltre a non pagarlo, bloccarlo, impedendogli di fare altri task (dello stesso batch o in futuro)
- Magari sulla base di analisi più sofisticate di quelle fatte online durante il task
- Ad es.:
 - O usando le risposte a tutti i task dello stesso worker
 - O analisi automatica offline dei testi
 - O analisi manuale dei testi e/o delle risposte
 - O usando le risposte degli altri worker

Accordo (e qualità)

- Assunzione, congettura:
- "Se i worker sono in accordo, la qualità è alta"
 - (disaccordo, bassa)
- Spesso è così
- Non sempre è così
 - Es.: nei nostri esperimenti di fact-checking, a volte alto accordo fra worker che danno risposta sbagliata

2 tipi di accordo

- Accordo come proxy / approssimazione della qualità
- Se non so il valore corretto di una risposta, posso comunque:
- 1) Guardare quanto i worker sono **in accordo fra di loro**, come insieme di tutti i worker che hanno svolto lo stesso task
 - Se i worker sono in accordo, **dati affidabili**
 - Se i worker sono in disaccordo, dati non affidabili
 - E, se dati non affidabili, magari raccoglierne ancora, cambiare, ecc.
- 2) Guardare quanto il singolo worker è **in accordo con gli altri worker** che hanno svolto lo stesso task
 - E assumere / ipotizzare che:
 - Se un worker è in accordo con gli altri, allora **qualità del work(er) alta**
 - Se un worker non è in accordo, allora qualità del work(er) bassa
 - E, al solito, escludere il work(er) con accordo/qualità bassi

2 tipi di accordo

- Accordo come proxy / approssimazione della qualità
- Se non so il valore corretto di una risposta, posso comunque:
- 1) Guardare quanto i worker sono **in accordo fra di loro**, come insieme di tutti i worker che hanno svolto lo stesso task
 - Se i worker sono in accordo, **dati affidabili**
 - Se i worker sono in disaccordo, dati non affidabili
 - E, se dati non affidabili, magari raccoglierne ancora, cambiare, ecc.
- 2) Guardare quanto il singolo worker è **in accordo con gli altri worker** che hanno svolto lo stesso task
 - E assumere / ipotizzare che:
 - Se un worker è in accordo con gli altri, allora **qualità del work(er)** alta
 - Se un worker non è in accordo, allora qualità del work(er) bassa
 - E, al solito, escludere il work(er) con accordo/qualità bassi
- Vediamo prima 1), poi 2)

Domanda

- Come definisco / misuro l'accordo (Agreement)?
 - E/o il disaccordo (Disagreement)
- "Quanto i vari worker che lavorano sullo stesso task danno la stessa risposta?"
 - "Quanto i vari worker che lavorano sullo stesso task danno risposta diverse?"
 - "(e quanto diverse?)"
- (non si usano Aggregazione e Ground Truth)

	t_1	t_2	t_3	...	t_n
w_1	v_{11}	v_{12}	v_{13}		v_{1n}
w_2	v_{21}		v_{23}	...	v_{2n}
w_3	v_{31}	v_{32}		...	v_{3n}
...				...	
w_m		v_{m2}	v_{m3}		

Aggr	av_1	av_2	av_3	...	av_n
------	--------	--------	--------	-----	--------

Truth	cv_1	cv_2	cv_3	...	cv_n
-------	--------	--------	--------	-----	--------

Domanda

- Come definisco / misuro l'accordo (Agreement)?
 - E/o il disaccordo (Disagreement)
- "Quanto i vari worker che lavorano sullo stesso task danno la stessa risposta?"
 - "Quanto i vari worker che lavorano sullo stesso task danno risposta diverse?"
 - "(e quanto diverse?)"
- (non si usano Aggregazione e Ground Truth)

	t_1	t_2	t_3	...	t_n
w_1	v_{11}	v_{12}	v_{13}		v_{1n}
w_2	v_{21}		v_{23}	...	v_{2n}
w_3	v_{31}	v_{32}		...	v_{3n}
...				...	
w_m		v_{m2}	v_{m3}		

Variabili

- # worker
- Tipo di risposte
 - Scala nominale
 - Binaria
 - Più valori
 - Ordinale
 - Intervalli
 - Rapporti
- Sparsità

(Tante!) Misure di accordo

- 2 worker
 - Percent(age) agreement
 - Cohen's kappa
 - Scott's Pi
- Più worker
 - Pairwise agreement
 - Fleiss's kappa
- Altre
 - Intraclass Correlation Coefficient (ICC)
 - Krippendorff's Alpha
 - Phi
 - ...

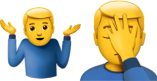
(Tante!) Misure di accordo

- 2 worker
 - Percent(age) agreement
 - Cohen's kappa
 - Scott's Pi
- Più worker
 - Pairwise agreement
 - Fleiss's kappa
- Altre
 - Intraclass Correlation Coefficient (ICC)
 - Krippendorff's Alpha
 - Phi
 - ...

(Terminologia)

- Reliability measures
 - "Affidabilità"
 - Di chi risponde, dei dati raccolti
- Raters
- Ratings
- Items

Nozione non semplice

- Varie possibilità
- Terminologia strana
- Mancanza di... accordo 
- Vediamo comunque qualcosa

Percent(age) agreement

- Accordo percentuale,
percentuale di accordo
- Partiamo dalla solita matrice
workers / tasks
- E la massaggiamo /
semplifichiamo

	t_1	t_2	t_3	...	t_n
w_1	v_{11}	v_{12}	v_{13}		v_{1n}
w_2	v_{21}		v_{23}	...	v_{2n}
w_3	v_{31}	v_{32}		...	v_{3n}
...				...	
w_m		v_{m2}	v_{m3}		

Caso semplice

- 2 worker
- Scala nominale binaria
 - Y (Yes) N (No)
- Non sparsa
 - Tutti (e due) i worker valutano tutto

	t_1	t_2	t_3	...	t_n
w_1	v_{11}	v_{12}	v_{13}		v_{1n}
w_2	v_{21}		v_{23}	...	v_{2n}

	t_1	t_2	t_3	...	t_n
w_1	Y	Y	Y		N
w_2	N		Y	N	N

	t_1	t_2	t_3	...	t_n
w_1	Y	Y	Y	Y	N
w_2	N	Y	Y	N	N

Rappresentazione alternativa

- Matrice workers / task \rightarrow "Confusion matrix"

	t_1	t_2	t_3	...	t_n
w_1	Y	Y	Y	Y	N
w_2	N	Y	Y	N	N

 \rightarrow

		w_2	
		Y	N
w_1	Y	2	2
	N	0	1

- Altro es. / ex.: scrivete la matrice worker / task...

		w_2	
		Y	N
w_1	Y	25	2
	N	1	18

Percent agreement (p)

- "Accordo percentuale"
- Con la matrice di confusione è facile da calcolare
- Percentuale in cui i due worker vanno d'accordo
 - Dicono la stessa cosa
- = Accuratezza, Accuracy
 - Considerando uno dei due (a caso) come ground truth, vedere quanto l'altro è d'accordo

$$p = \frac{a+d}{a+b+c+d}$$

		w_2	
		Y	N
w_1	Y	a	b
	N	c	d

Pro

- Facile da calcolare
 - $(25+35)/(25+20+20+35) = 0.6$
- Intuitivo
 - w_1 e w_2 sono in accordo nel 60% dei casi
- Consente confronti
 - w_1 e w_2 sono più in accordo di w_3 e w_4 (0.4)
- Estendibile facilmente a **più categorie**
 - Si aggiungono righe e colonne
 - Si conta sempre la diagonale
 - $(70 / 170)$

		w_2	
		Y	N
w_1	Y	25	20
	N	20	35

		w_4	
		Y	N
w_3	Y	20	35
	N	25	20

		w_6		
		Y	N	?
w_5	Y	20	35	15
	N	25	20	10
	?	15	25	30

Contro (1/2)

- Agreement by chance (accordo casuale)
- 2 Worker che sparano a caso
- Con **2** categorie (Y, N)
 - Accordo del 50%!? ($1/4 + 1/4$)
- Con **3** categorie
 - Accordo del 33% ($1/9 + 1/9 + 1/9$)
- Con **4** categorie
 - Accordo del 25% ($1/16 + 1/16 + 1/16 + 1/16$)
- Con **K** categorie
 - Esercizio

		w_2	
		Y	N
w_1	Y	25	25
	N	25	25

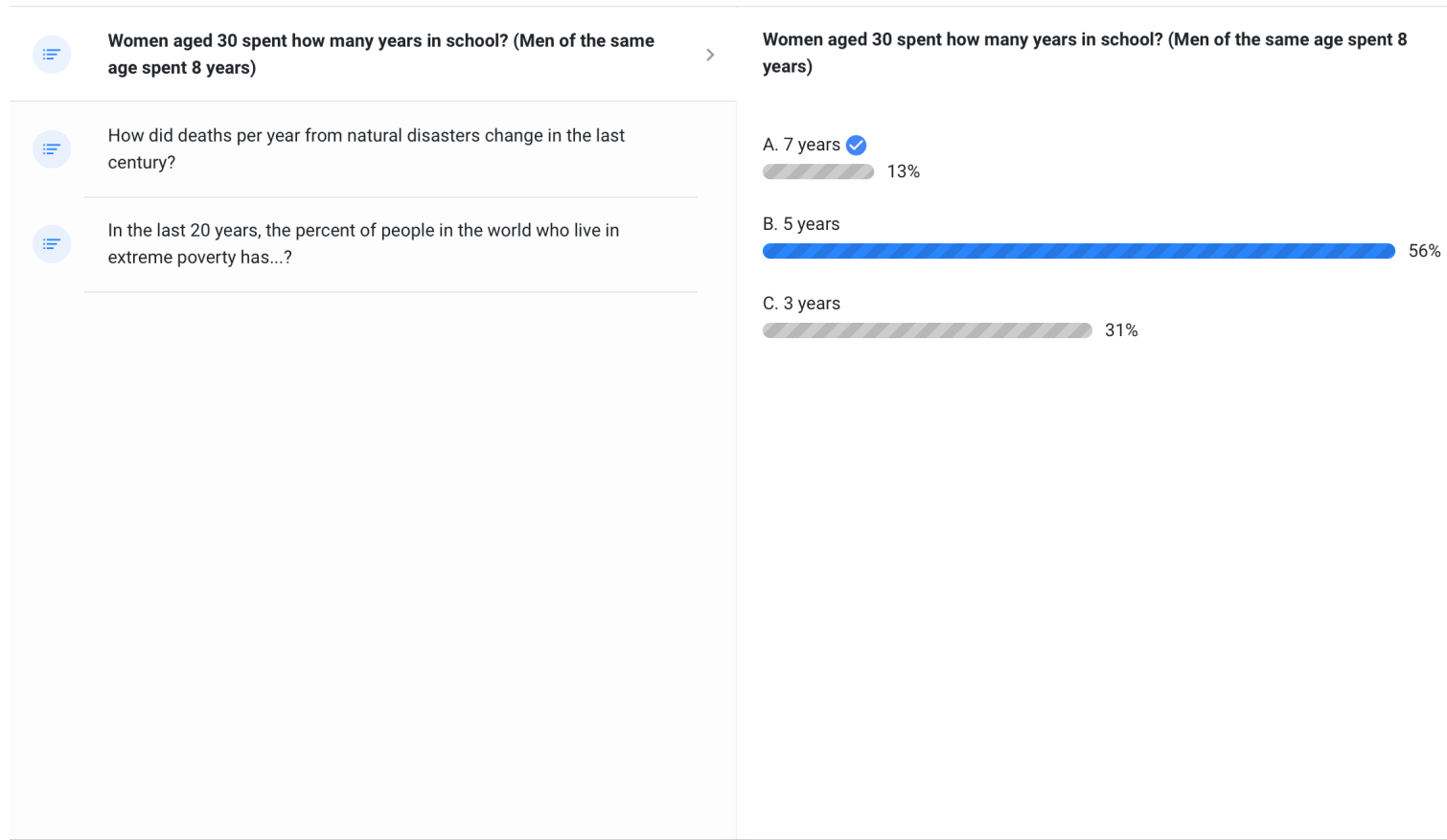
		w_4		
		Y	N	?
w_3	Y	33	33	33
	N	33	33	33
	?	33	33	33

		w_6			
		A	B	C	D
w_5	A	6	6	6	6
	B	6	6	6	6
	C	6	6	6	6
	D	6	6	6	6

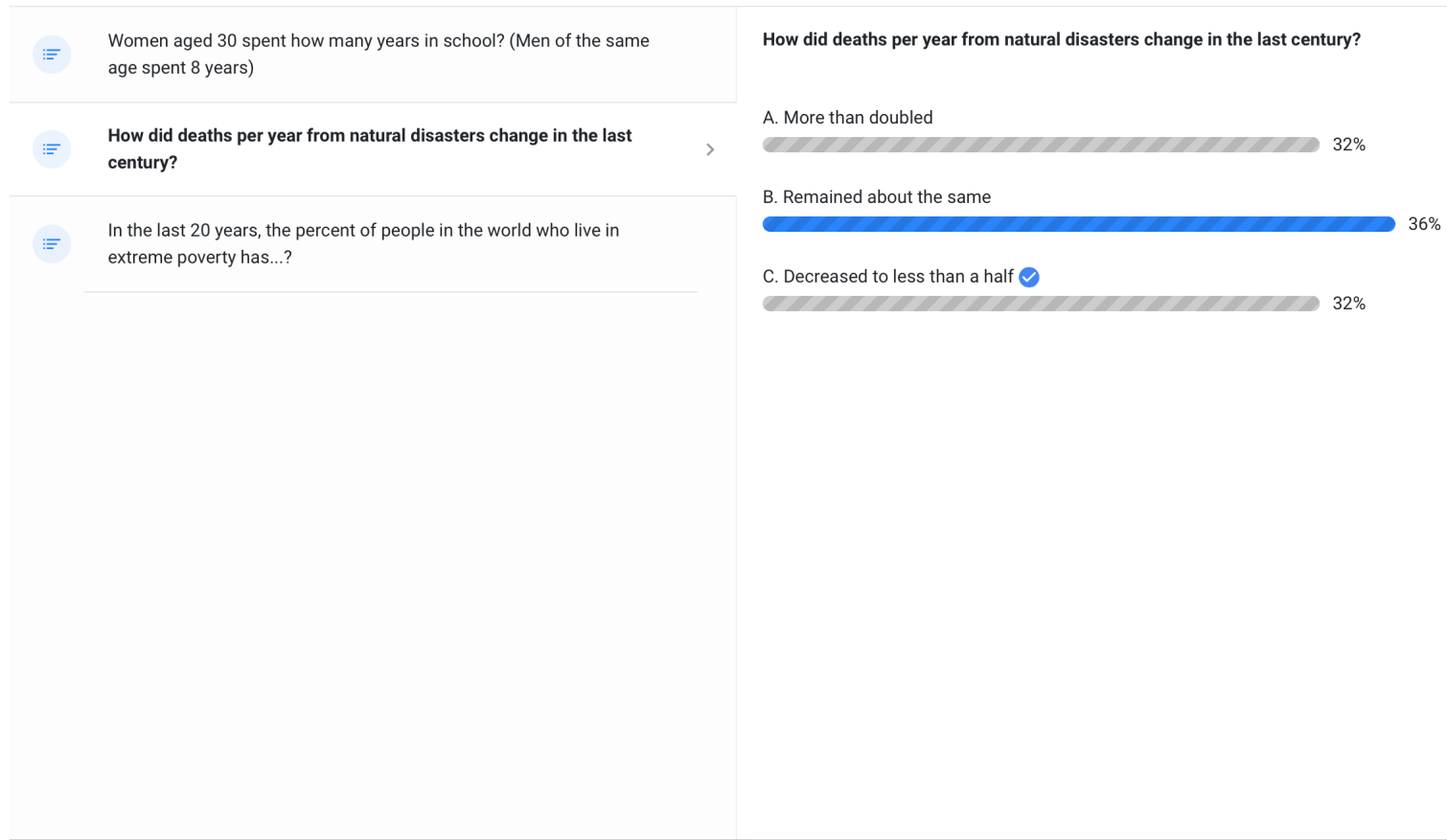
Sull'agreement by chance (e altro)

- https://www.ted.com/talks/hans_and_ola_rosling_how_not_to_be_ignorant_about_the_world/transcript
- <https://www.sli.do/> #T987, #T988, #T989
- Compiti per casa (entro 14/12):
 - 1) Guardare il video per intero, ma...
 - 2) ... rispondendo alle 3 domande su slido man mano
 - 3) Poi, riflettere sull'agreement by chance
- Gli highlight secondo me:
 - "So I went to the zoo and I asked the chimps"
 - "Aaahhhh, you almost made it to the chimps"

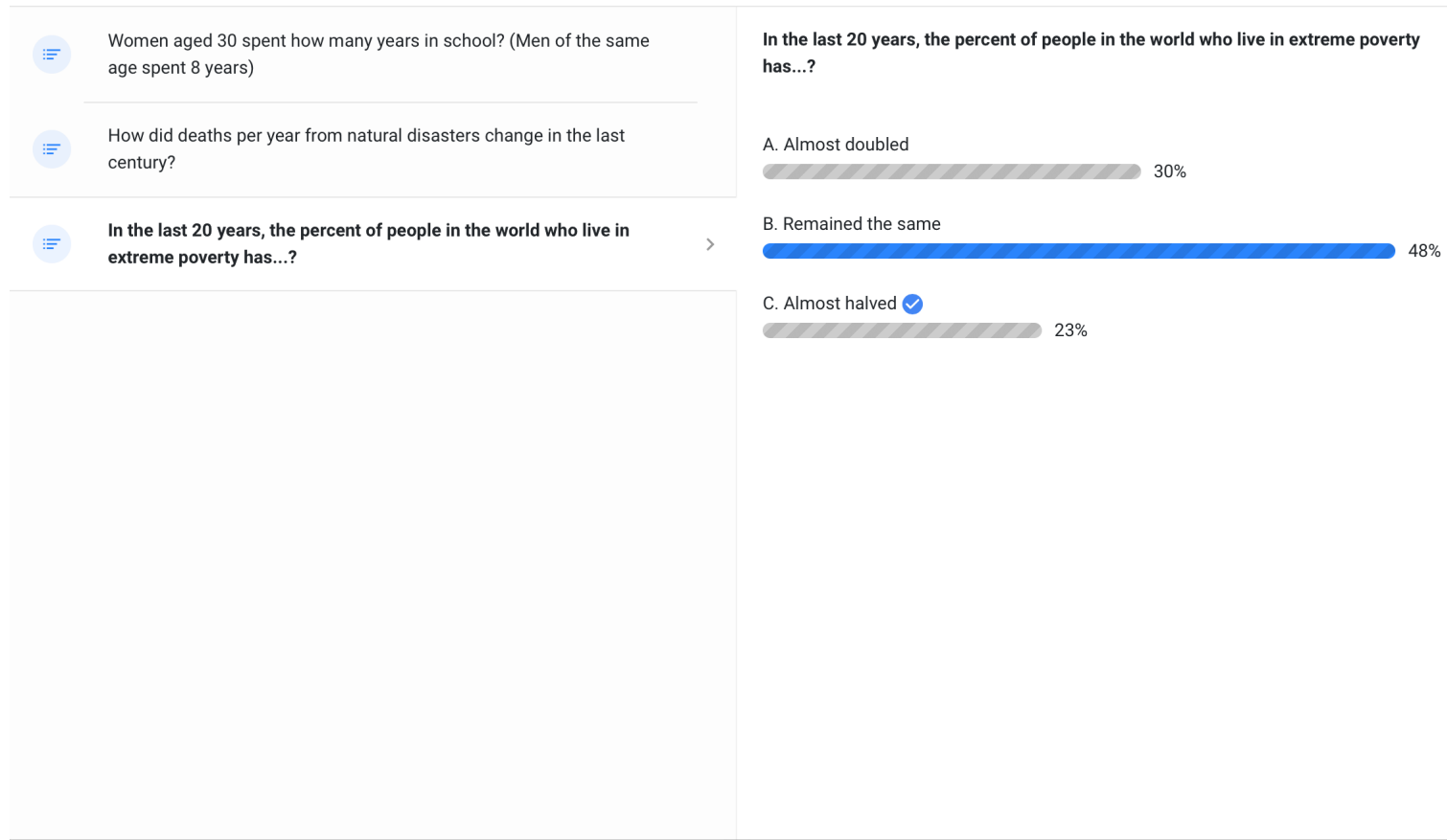
Le risposte dell'a.a. 19/20 (1/3) 🙋🧐



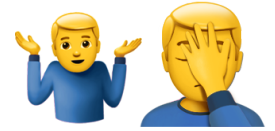
Le risposte dell'a.a. 19/20 (2/3) 🙋🧐



Le risposte dell'a.a. 19/20 (3/3) 🙋🧐



Le risposte dell'a.a. 20/21 (1/3)



How did deaths per year from natural disasters change in the last century?

11  >

How did deaths per year from natural disasters change in the last century?

A. More than doubled

 27%

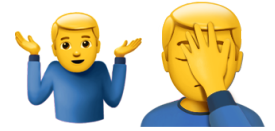
B. Remained about the same

 55%

C. Decreased to less than a half 

 18%


Le risposte dell'a.a. 20/21 (2/3)



Women aged 30 spent how many years in school?
(Men of the same age spent 8 years)

11  >

Women aged 30 spent how many years in school? (Men of the same age spent 8 years)

A. 7 years 



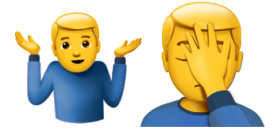
B. 5 years



C. 3 years



Le risposte dell'a.a. 20/21 (3/3)



In the last 20 years, the percent of people in the world who live in extreme poverty has...?

10  >

In the last 20 years, the percent of people in the world who live in extreme poverty has...?

A. Almost doubled



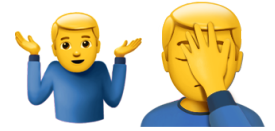
B. Remained the same



C. Almost halved 

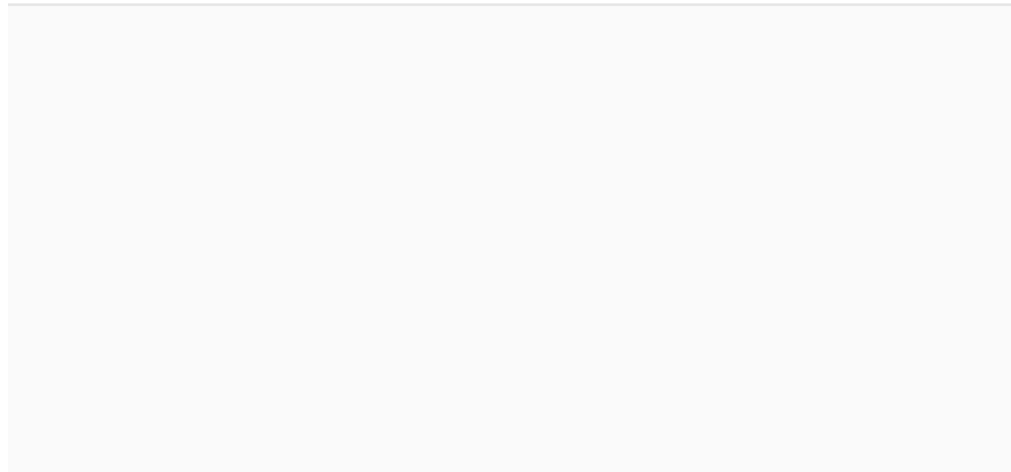


Le risposte dell'a.a. 21/22 (1/3)



How did deaths per year from natural disasters change in the last century?

3  >



How did deaths per year from natural disasters change in the last century?

A. More than doubled



B. Remained about the same



C. Decreased to less than a half 



Le risposte dell'a.a. 21/22 (2/3) 🙋🧐



Women aged 30 spent how many years in school? (Men of the same age spent 8 3 👤 > years)

Women aged 30 spent how many years in school? (Men of the same age spent 8 years)

A. 7 years ✓

33%

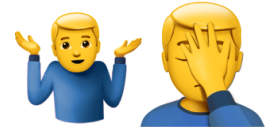
B. 5 years

67%

C. 3 years

0%

Le risposte dell'a.a. 21/22 (3/3)



In the last 20 years, the percent of people in the world who live in extreme poverty has...? 3 👤 >

In the last 20 years, the percent of people in the world who live in extreme poverty has...?

A. Almost doubled

 33%

B. Remained the same

 67%

C. Almost halved ✓

 0%

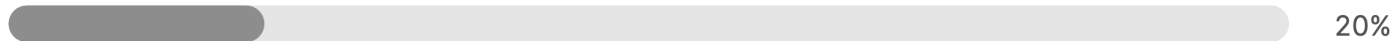
Le vostre risposte (1/3)



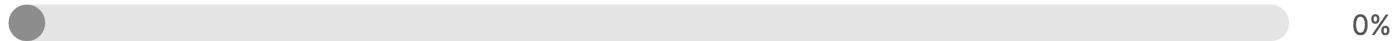
How did deaths per year from natural disasters change in the last century?

Multiple Choice Poll ☒ 5 votes  5 participants

A. More than doubled - 1 vote



B. Remained about the same - 0 votes



C. Decreased to less than a half - 4 votes ☒



Le vostre risposte (2/3)



Women aged 30 spent how many years in school? (Men of the same age spent 8 years)

Multiple Choice Poll ☒ 5 votes  5 participants

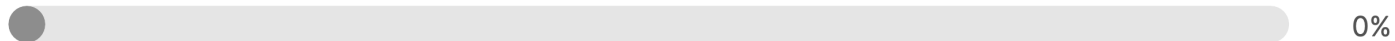
A. 7 years - 5 votes ☒



B. 5 years - 0 votes



C. 3 years - 0 votes



Le vostre risposte (3/3)



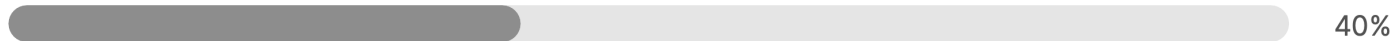
In the last 20 years, the percent of people in the world who live in extreme poverty has...?

Multiple Choice Poll ☒ 5 votes  5 participants

A. Almost doubled - 1 vote



B. Remained the same - 2 votes



C. Almost halved - 2 votes ☒



Contro (2/2)

- Categorie sbilanciate
- Se i worker sanno che il 90% delle risposte sono N, cosa faranno?
- Diranno al 90% N!
 - O magari N al 100%
- Accordo del 90%!? $(85+5)/100$
 - O del 100%
 - (E alta accuracy dei singoli worker con la ground truth)
 - (dicendo N non sbaglio 9 volte su 10...)
- (più worker, $m > 2$ worker?)

		w_2		
		Y	N	
w_1	Y	25	25	10
	N	25	25	

10 90

		w_2		
		Y	N	
w_1	Y	0	0	
	N	0	100	

Cohen's kappa (κ)

- Idea: Cercare di migliorare/correggere il percent agreement tenendo conto dell'agreement by chance
- Sempre 2 worker
- Sempre scala nominale
- $$\kappa = \frac{p_o - p_e}{1 - p_e}$$
- p_o è il percent agreement osservato
 - (o = observed)
- p_e è il percent agreement by chance, atteso
 - (e = expected)
 - Calcolato usando i dati per stimare le probabilità che ogni worker scelga ogni data categoria
 - (se zero, = a prima)
 - (se uno, mi aspetto accordo totale anche rispondendo a caso)

Esempio (1/3)

- 50 task
 - Ad es., dire per 50 foto se contengono un'auto o no
- 2 worker
 - Ogni foto viene giudicata da entrambi (no sparsità)

	t_1	t_2	t_3	...	t_{50}
w_1	Y	Y	Y	...	N
w_2	N	Y	Y	...	N



		w_2	
		Y	N
w_1	Y	20	5
	N	10	15

- $p_o = \frac{20+15}{20+5+10+15} = \frac{35}{50} = 0.7$ (come prima)

Esempio (2/3)

		w_2		
		Y	N	
w_1	Y	a	b	$a + b$
	N	c	d	$c + d$
		$a + c$	$b + d$	$a + b + c + d$

- $E p_e$?
- w_1 ha detto 25 volte Y (e 25 volte N)
 - w_1 ha detto 50% di volte Y (e 50% di volte N)
- w_2 ha detto 30 volte Y (e 20 volte N)
 - w_2 ha detto 60% di volte Y (e 40% di volte N)
- Probabilità che entrambi dicano Y
 - $p_Y = \frac{a+b}{a+b+c+d} \cdot \frac{a+c}{a+b+c+d} = 0.5 \times 0.6 = 0.3$
- Probabilità che entrambi dicano N
 - $p_N = \frac{c+d}{a+b+c+d} \cdot \frac{b+d}{a+b+c+d} = 0.5 \times 0.4 = 0.2$
- $p_e = p_Y + p_N = 0.3 + 0.2 = 0.5$

		w_2		
		Y	N	
w_1	Y	20	5	25
	N	10	15	25
		30	20	50

Esempio (3/3)

- E quindi
- $\kappa = \frac{p_o - p_e}{1 - p_e} = \frac{0.7 - 0.5}{1 - 0.5} = 0.4$
- Notare che κ può essere < 0
 - Peggior dell'accordo casuale (!)

Scott's Pi (π)

- Stessa idea di Cohen's kappa κ
 - Correggere / normalizzare sulla base dell'accordo casuale
 - Accordo casuale calcolato in altro modo
- Non la vediamo
- Limitazioni di Cohen's kappa (κ) e Scott's Pi (π)
 - Dati categoriali
 - Due soli worker

(Tante!) Misure di accordo

- 2 worker
 - Percent(age) agreement
 - Cohen's kappa
 - Scott's Pi
- Più worker
 - Pairwise agreement
 - Fleiss's kappa
- Altre
 - Intraclass Correlation Coefficient (ICC)
 - Krippendorff's Alpha
 - Phi
 - ...

Fleiss kappa (κ)

- Estende Scott's Pi (π) a m worker
- Sempre scala nominale
- Sempre # categorie a piacere

	t_1	t_2	t_3	...	t_n
w_1	A	B	D	E	A
w_2	B	B	D	D	A
...
w_m	A	B	C	D	A

Prima domanda / problema

- Come misuro l'accordo su m worker?!?
 - AAAAA accordo completo
 - AAAAB>AAABB
 - ...sembra che potrei usare la max frazione di worker che danno la stessa risposta
- Ma:
 - AAABB>AAABC?!
 - Sempre 3/5 ma più accordo nel primo caso!
- Soluzione:
 - Pairwise agreement, Accordo a coppie
 - Considero tutte le coppie di worker e vedo quale frazione di coppie è in accordo

	t_1	t_2	t_3	...	t_n
w_1	A	B	D	E	A
w_2	B	B	D	D	A
...
w_m	A	B	C	D	A

Pairwise agreement

- Si confrontano non i singoli valori espressi da un worker, ma le coppie di valori
 - (OK per scale N, O, I, R)
- m worker, $\binom{m}{2} = m(m-1)/2$ coppie di worker
- % di coppie in accordo (disaccordo)
- Es. (5 task, 4 worker, 3 categorie ABC)
 - 4 worker, $4 \times 3 / 2 = 6$ coppie:
 - $(w_1, w_2) (w_1, w_3) (w_1, w_4)$
 $(w_2, w_3) (w_2, w_4) (w_3, w_4)$

	t_1	t_2	t_3	t_4	t_5
Pairwise agreement	6/6	3/6	2/6	1/6	6/6

	t_1	t_2	t_3	t_4	t_5
w_1	A	A	A	A	B
w_2	A	A	A	A	B
w_3	A	A	B	B	B
w_4	A	B	B	C	B

Fleiss kappa (κ) def. (1/3)

- Definizione generale:

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e}$$

- ~ Cohen's kappa, correggo per accordo casuale
- Però con pairwise agreement invece di percent agreement:
 - \bar{P} è il pairwise agreement osservato
 - \bar{P}_e è il pairwise agreement atteso, casuale
- (Aspetti tecnici ed esempio, non li vediamo)

Fleiss kappa (κ) def. (2/3)

- Notazione
 - m : numero di worker
 - n : numero di item (task)
 - k : numero di categorie
 - n_{ij} : numero di worker che hanno assegnato l'item i alla categoria j
- Calcoliamo la proporzione di assegnamenti a ogni categoria j
 - $p_j = \frac{1}{nm} \sum_{i=1}^n n_{ij}$ ($1 = \sum_{j=1}^k p_j$)
- E poi la somma dei loro quadrati, e questo è \bar{P}_e
 - $\bar{P}_e = \sum_{j=1}^k p_j^2$
- $p_j \leq 1 \rightarrow$
 - valore max. di \bar{P}_e : tutti gli assegnamenti a una categoria
 - valore min. di \bar{P}_e : tutti gli assegnamenti equidistribuiti

Fleiss kappa (κ) def. (3/3)

- Calcoliamo la misura di quanto i worker sono in accordo sull'i-esimo item (pairwise agreement)

- Quante **coppie worker-worker sono in accordo**, rispetto a **tutte le coppie possibili**

$$\begin{aligned}
 P_i &= \frac{2}{m(m-1)} \sum_{j=1}^k \frac{n_{ij}(n_{ij}-1)}{2} = \\
 &= \frac{1}{m(m-1)} \sum_{j=1}^k (n_{ij}^2 - n_{ij}) = \\
 &= \frac{1}{m(m-1)} [(\sum_{j=1}^k n_{ij}^2) - n]
 \end{aligned}$$

- E poi la media su tutti gli item

$$\begin{aligned}
 \bar{P} &= \frac{1}{n} \sum_{i=1}^n P_i = \frac{1}{n} \sum_{i=1}^n \left(\frac{1}{m(m-1)} [(\sum_{j=1}^k n_{ij}^2) - n] \right) = \\
 &= \frac{1}{nm(m-1)} (\sum_{i=1}^n \sum_{j=1}^k n_{ij}^2 - nm)
 \end{aligned}$$

Esempio

- $m = 14$ worker
 - $w_1 w_2 \dots w_{14}$
- Assegnano $n = 10$ item
 - $i_1 i_2 \dots i_{10}$
- ($m \times n = 14 \times 10 = 140$ assegnamenti)
- a $k = 5$ categorie
 - ABCDE
 - N.B. **Indipendenti, NON ordinate**

	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8	i_9	i_{10}
w_1	E	B	C	B	A	A	A	A	A	B
w_2	E	B	C	B	A	A	A	A	A	B
w_3	E	C	C	B	B	A	A	B	A	C
w_4	E	C	D	C	B	A	B	B	A	C
w_5	E	C	D	C	C	A	B	B	A	D
w_6	E	C	D	C	C	A	C	B	A	D
w_7	E	C	D	C	C	A	C	B	B	D
w_8	E	C	D	C	C	B	C	C	B	E
w_9	E	D	E	C	C	B	C	C	B	E
w_{10}	E	D	E	C	C	B	C	C	B	E
w_{11}	E	D	E	C	C	B	C	D	B	E
w_{12}	E	D	E	C	C	B	D	D	C	E
w_{13}	E	E	E	D	D	B	D	E	C	E
w_{14}	E	E	E	D	E	B	D	E	D	E

Esempio

- ➡ Altra rappresentazione, con n_{ij} espliciti
- Agreement table
 - Non è la matrice worker / task!
 - Non è la confusion matrix!!
 - "Matrice categorie / task"

	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8	i_9	i_{10}
A	0	0	0	0	2	7	3	2	6	0
B	0	2	0	3	2	7	2	5	5	2
C	0	6	3	9	8	0	6	3	2	2
D	0	4	5	2	1	0	3	2	1	3
E	14	2	6	0	1	0	0	2	0	7

Esempio

- Calcoliamo Fleiss kappa (κ)

	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8	i_9	i_{10}	Tot
A	0	0	0	0	2	7	3	2	6	0	20
B	0	2	0	3	2	7	2	5	5	2	28
C	0	6	3	9	8	0	6	3	2	2	39
D	0	4	5	2	1	0	3	2	1	3	21
E	14	2	6	0	1	0	0	2	0	7	32

140

Esempio

- Calcoliamo Fleiss kappa (κ)
- p_j è la proporzione di assegnamenti (su 140) nella categoria j

	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8	i_9	i_{10}	Tot	p_j
A	0	0	0	0	2	7	3	2	6	0	20	.143
B	0	2	0	3	2	7	2	5	5	2	28	.200
C	0	6	3	9	8	0	6	3	2	2	39	.279
D	0	4	5	2	1	0	3	2	1	3	21	.150
E	14	2	6	0	1	0	0	2	0	7	32	.229

140

Esempio

- Calcoliamo Fleiss kappa (κ)
- p_j è la proporzione di assegnamenti (su 140) nella categoria j
- P_i è il pairwise agreement sull'item i

	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8	i_9	i_{10}	Tot	p_j
A	0	0	0	0	2	7	3	2	6	0	20	.143
B	0	2	0	3	2	7	2	5	5	2	28	.200
C	0	6	3	9	8	0	6	3	2	2	39	.279
D	0	4	5	2	1	0	3	2	1	3	21	.150
E	14	2	6	0	1	0	0	2	0	7	32	.229
P_i	1.00	.253	.308	.440	.330	.462	.242	.176	.286	.286	140	

Conti

- Media dei P_i :

- $\bar{P} = \frac{1}{n} \sum_{i=1}^n P_i = \frac{1}{10} \sum_{i=1}^{10} P_i =$
 $= \frac{1}{10} (1.000 + 0.253 + \dots + 0.286) = 0.378$

- Somma dei quadrati dei p_j :

- $\bar{P}_e = 0.143^2 + 0.200^2 + 0.279^2 + 0.150^2 + 0.229^2 = 0.213$

- $\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e} = \frac{0.378 - 0.213}{1 - 0.213} = 0.210$

Interpretazione (controversa!)

Value	Interpretation
< 0	Poor agreement
0.01 – 0.20	Slight agreement
0.21 – 0.40	Fair agreement
0.41 – 0.60	Moderate agreement
0.61 – 0.80	Substantial agreement
0.81 – 1.00	Almost perfect agreement

ICC

(Intraclass Correlation Coefficient)

Krippendorff's Alpha

Phi

Solo cenni...

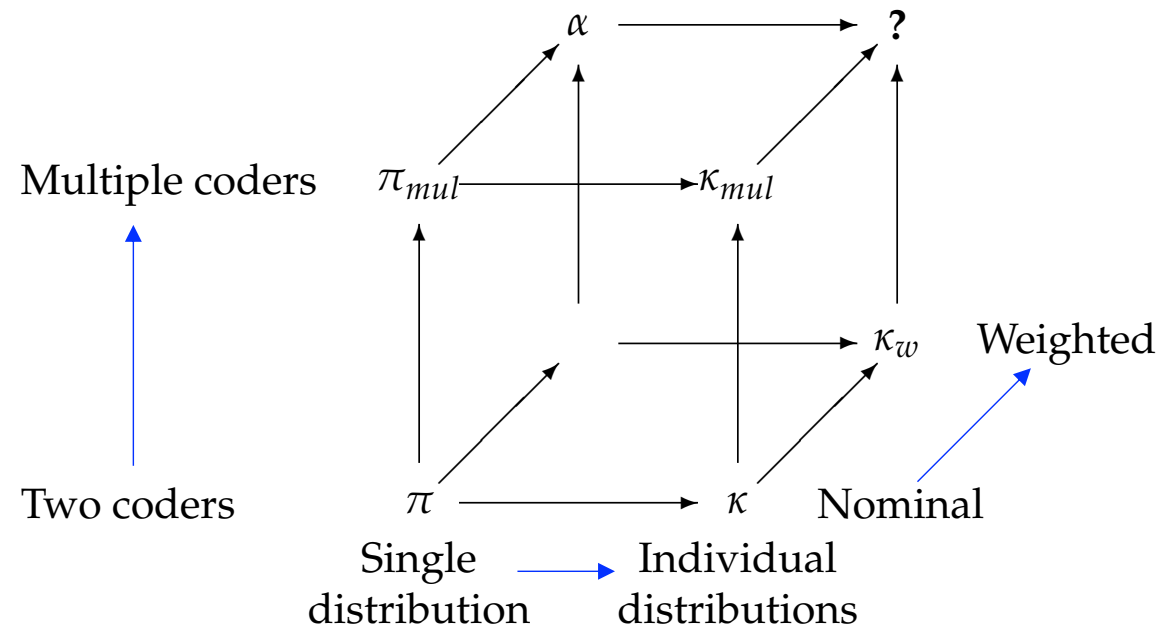


Figure 1
Generalizing π along three dimensions

[Artstein & Poesio, 2007]

(Tante!) Misure di accordo

- 2 worker
 - Percent(age) agreement
 - Cohen's kappa
 - Scott's Pi
- Più worker
 - Pairwise agreement
 - Fleiss's kappa
- Altre
 - Intraclass Correlation Coefficient (ICC)
 - Krippendorff's Alpha
 - Phi
 - ...

L'accordo, di nuovo

- Accordo come proxy / approssimazione della qualità
- Se non so il valore corretto di una risposta, posso comunque:
- 1) Guardare quanto i worker sono **in accordo fra di loro**, come insieme di tutti i worker che hanno svolto lo stesso task
 - Se i worker sono in accordo, **dati affidabili**
 - Se i worker sono in disaccordo, dati non affidabili
 - E, se dati non affidabili, magari raccoglierne ancora, modifiche, ecc.
- 2) Guardare quanto il singolo worker è **in accordo con gli altri worker** che hanno svolto lo stesso task
 - E assumere / ipotizzare che:
 - Se un worker è in accordo con gli altri, allora **qualità del work(er)** alta
 - Se un worker non è in accordo, allora qualità del work(er) bassa
 - E, al solito, escludere il work(er) con accordo/qualità bassi

Accordo di un worker con l'aggregazione

- Non guardo l'agreement generale di un insieme di worker ma:
 - considero ogni singolo worker
 - guardo quanto correla con l'aggregato degli altri worker (sullo stesso task)

	t_1	t_2	t_3	...	t_n
w_1	v_{11}	v_{12}	v_{13}		v_{1n}
w_2	v_{21}		v_{23}	...	v_{2n}
w_3	v_{31}	v_{32}		...	v_{3n}
...				...	
w_m		v_{m2}	v_{m3}		

Aggr	av_1	av_2	av_3	...	av_n
------	--------	--------	--------	-----	--------

Truth	cv_1	cv_2	cv_3	...	cv_n
-------	--------	--------	--------	-----	--------

Accordo di un worker con l'aggregazione

- Ritorniamo al caso non sparso

	t_1	t_2	t_3	...	t_n
w_1	v_{11}	v_{12}	v_{13}		v_{1n}
w_2	v_{21}		v_{23}	...	v_{2n}
w_3	v_{31}	v_{32}		...	v_{3n}
...				...	
w_m		v_{m2}	v_{m3}		

Aggr	av_1	av_2	av_3	...	av_n
------	--------	--------	--------	-----	--------

Truth	cv_1	cv_2	cv_3	...	cv_n
-------	--------	--------	--------	-----	--------

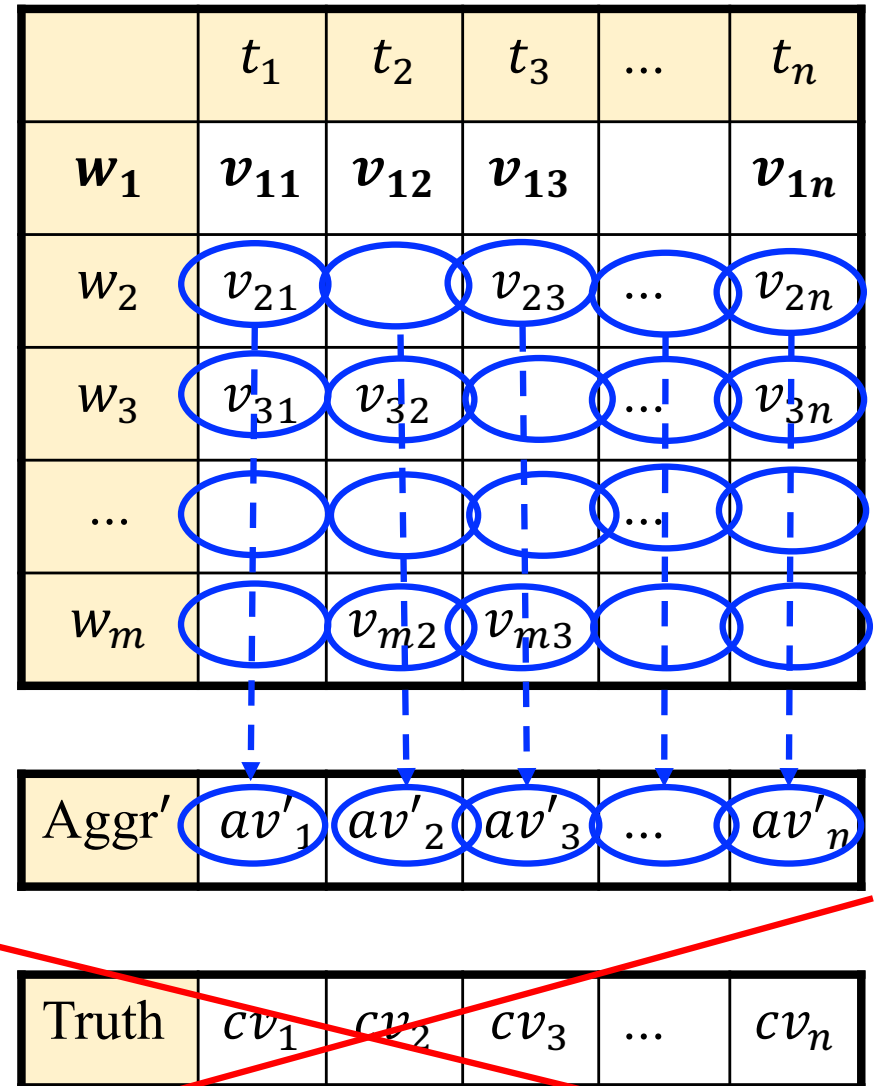
Accordo di un worker con l'aggregazione

- Ad es., w_1
 - Voglio sapere quanto è in accordo con la media degli altri worker
 - È una misura di qualità di w_1

	t_1	t_2	t_3	...	t_n
w_1	v_{11}	v_{12}	v_{13}		v_{1n}
w_2	v_{21}		v_{23}	...	v_{2n}
w_3	v_{31}	v_{32}		...	v_{3n}
...				...	
w_m		v_{m2}	v_{m3}		

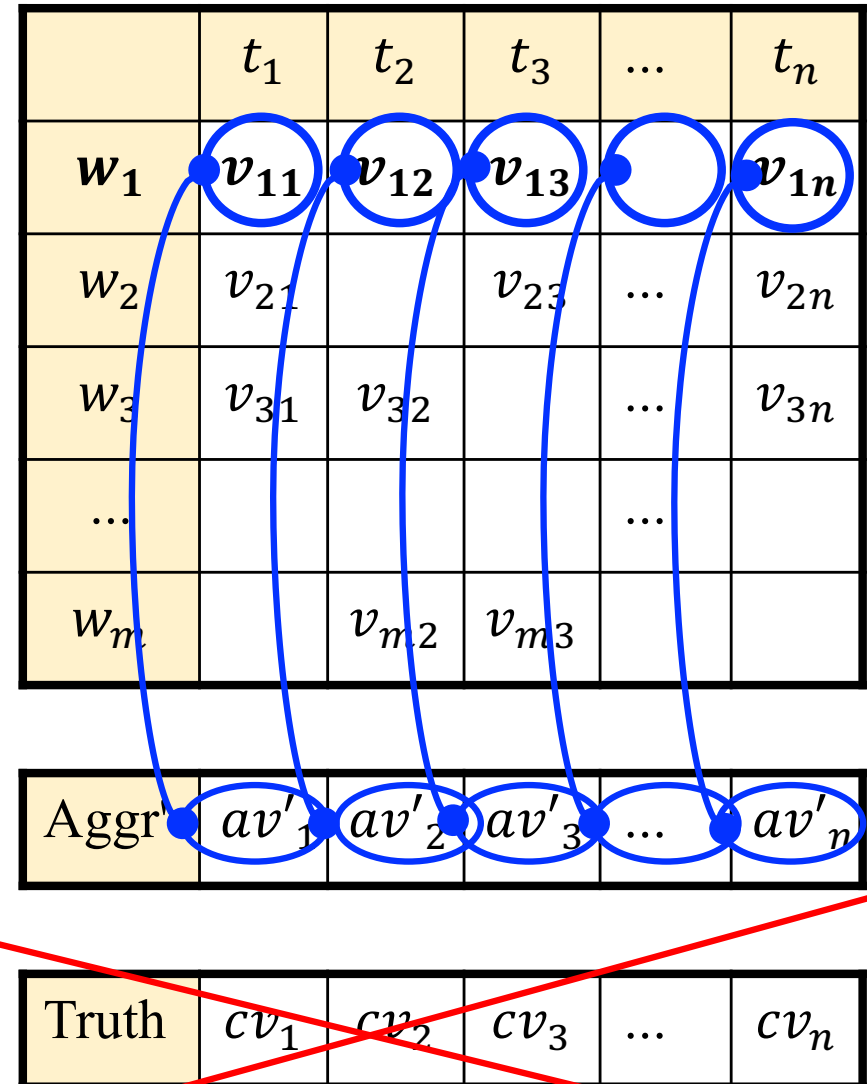
Accordo di un worker con l'aggregazione

- Ad es., w_1
 - Voglio sapere quanto è in accordo con la media degli altri worker
 - È una misura di qualità di w_1
- Aggrego tutti **meno** w_1
- N.B.
 $\text{Aggr}' = [av'_1, av'_2, av'_3, \dots, av'_n]$
 non è il valore aggregato "vero"
 (ma cambia poco)



Accordo di un worker con l'aggregazione

- Ad es., w_1
 - Voglio sapere quanto è in accordo con la media degli altri worker
 - È una misura di qualità di w_1
- Aggrego tutti **meno** w_1
- N.B.
 $\text{Aggr}' = [av'_1, av'_2, av'_3, \dots, av'_n]$
 non è il valore aggregato "vero"
 (ma cambia poco)
- Misuro agreement fra valori di w_1 e aggregazioni
- Lo faccio per tutti i w_i e ho una loro misura di qualità



E poi? Come sfrutto l'accordo?

- (nel senso di misura di qualità per singolo worker)
- Ad esempio escludendo i worker che hanno agreement con l'aggregazione (~qualità) inferiore
 - Magari iterativamente
 - Prima 1, il peggiore
 - Poi 2, anche il secondo peggiore
 - ...
- O pesarli di meno nell'aggregazione
 - Media pesata
- O con un modello più sofisticato
- "Co-determination algorithm"
 - Si "determinano insieme" l'aggregazione e la qualità dei worker

Framework [Li et al.]

Algorithm 1 Solution framework

Input: workers' answers V

Output: inferred truth v_i^* ($1 \leq i \leq n$), worker quality q^w ($w \in \mathcal{W}$)

```
1: Initialize all workers' qualities ( $q^w$  for  $w \in \mathcal{W}$ );
2: while true do
3:   // Step 1: Inferring the Truth
4:   for  $1 \leq i \leq n$  do
5:     Inferring the truth  $v_i^*$  based on  $V$  and  $\{q^w \mid w \in \mathcal{W}\}$ ;
6:   end for
7:   // Step 2: Estimating Worker Quality
8:   for  $w \in \mathcal{W}$  do
9:     Estimating the quality  $q^w$  based on  $V$  and  $\{v_i^* \mid 1 \leq i \leq n\}$ ;
10:  end for
11:  // Check for Convergence
12:  if Converged then
13:    break;
14:  end if
15: end while
16: return  $v_i^*$  for  $1 \leq i \leq n$  and  $q^w$  for  $w \in \mathcal{W}$ ;
```

Framework [Li et al.]

Algorithm 1 Solution framework

Input: workers' answers V

Output: inferred truth v_i^* ($1 \leq i \leq n$), worker quality q^w ($w \in \mathcal{W}$)

```
1: Initialize all workers' qualities ( $q^w$  for  $w \in \mathcal{W}$ );  
2: while true do  
3:   // Step 1: Inferring the Truth  
4:   for  $1 \leq i \leq n$  do  
5:     Inferring the truth  $v_i^*$  based on  $V$  and  $\{q^w \mid w \in \mathcal{W}\}$ ;  
6:   end for  
7:   // Step 2: Estimating Worker Quality  
8:   for  $w \in \mathcal{W}$  do  
9:     Estimating the quality  $q^w$  based on  $V$  and  $\{v_i^* \mid 1 \leq i \leq n\}$ ;  
10:  end for  
11:  // Check for Convergence  
12:  if Converged then  
13:    break;  
14:  end if  
15: end while  
16: return  $v_i^*$  for  $1 \leq i \leq n$  and  $q^w$  for  $w \in \mathcal{W}$ ;
```

Worker quality iniziale = per tutti, o "storica"

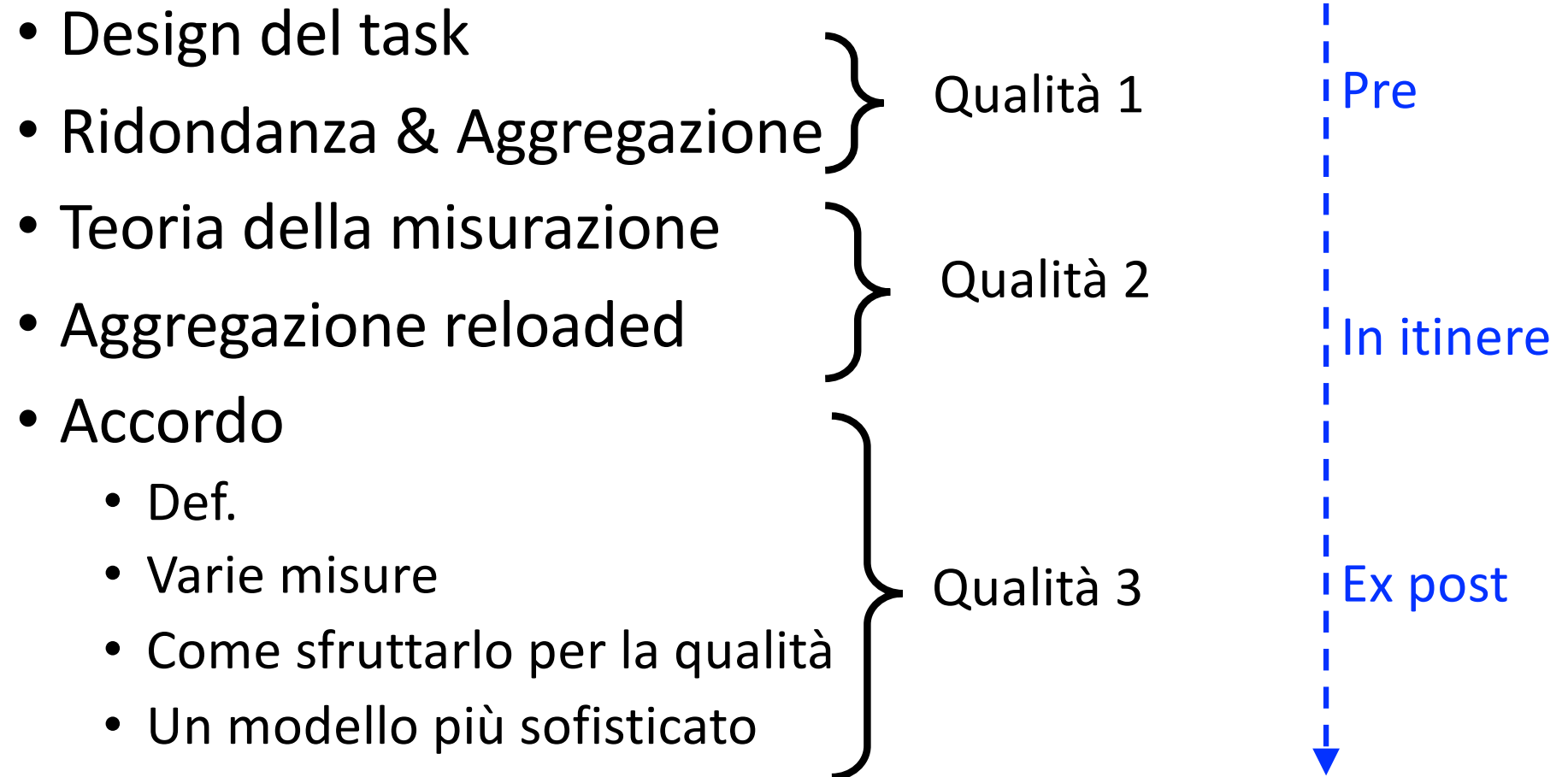
Aggregazione: media pesata con la qualità

Qualità = accordo con valore aggregato

Riassunto

- Continuare a discutere della qualità (parte 3)
- Ex post
- (Dis)Accordo – (Dis)Agreement
 - Nozione importante
 - Misurabile in vari modi 🙋
- Sfruttare l'accordo come proxy per la qualità
 - 1. Per l'affidabilità dei dati raccolti, capire se ho "buoni dati"
 - Varie misure di accordo
 - Def. Intuitive. Non viste tutte; tecnicismi...
 - 2. Per la qualità del singolo worker
 - "I worker più in disaccordo con gli altri sono di qualità inferiore"
 - Un modello più sofisticato

Schema



Nota

- Nel caso soggettivo
 - "cosa ti piace", non
 - "cosa è giusto"
- L'accordo non è una buona misura

	t_1	t_2	t_3	...	t_n
w_1	v_{11}	v_{12}	v_{13}		v_{1n}
w_2	v_{21}		v_{23}	...	v_{2n}
w_3	v_{31}	v_{32}		...	v_{3n}
...				...	
w_m		v_{m2}	v_{m3}		

Truth	cv_1	cv_2	cv_3	...	cv_n
-------	--------	--------	--------	-----	--------

Nota (un'altra)

- Non vi ho dato la **Soluzione Finale Da Usare In Tutti I Casi**
- Perché non esiste
 - (o almeno non la conosco!)
- Avete un armamentario da usare
 - In pratica, per attività concrete
 - Concettualmente, per comprendere altre soluzioni che potrebbero arrivare in futuro

Biblio

- https://en.wikipedia.org/wiki/Cohen%27s_kappa
- https://en.wikipedia.org/wiki/Scott%27s_Pi
- https://en.wikipedia.org/wiki/Fleiss%27_kappa
- Non viste:
 - ICC: https://en.wikipedia.org/wiki/Intraclass_correlation
 - Alpha:
https://en.wikipedia.org/wiki/Krippendorff%27s_alpha
 - Phi: (!)
<https://aaai.org/ocs/index.php/HCOMP/HCOMP17/paper/view/15927>

[Artstein & Poesio, 2007]

- Artstein & Poesio, 2007.
Inter-Coder Agreement for Computational Linguistics,
Journal of Computational Linguistics
- Per approfondimenti
- <http://ict.usc.edu/pubs/Inter-Coder%20Agreement%20for%20Computational%20Linguistics.pdf>

Survey Article

A shortened version of this article was submitted to the journal Computational Linguistics; this is the full version.

Inter-Coder Agreement for Computational Linguistics

Ron Artstein
University of Essex

Massimo Poesio
Università di Trento / University of
Essex

This article is a survey of issues concerning the measurement of agreement among corpus annotators. It exposes the mathematics and underlying assumptions of agreement coefficients such as Cohen's kappa and Krippendorff's alpha; relates these coefficients to explicit models of annotator error; discusses the use of coefficients in several annotation tasks; and argues that weighted alpha-like coefficients, traditionally less used than kappa-like measures in Computational Linguistics, may be more appropriate for many corpus annotation tasks – but that their use makes the interpretation of the value of the coefficient even harder.

1. Introduction and Motivations

Ever since the mid-Nineties, increasing effort has gone into putting semantics and discourse research on the same empirical footing as other areas of Computational Linguistics (CL). This soon led to worries about the subjectivity of the judgments required to create annotated resources, much greater for semantics and pragmatics than for the aspects of language interpretation of concern to the first resource creation efforts such as the creation of the Brown corpus (Francis and Kucera 1982), the British National Corpus (Leech, Garside, and Bryant 1994) or the Penn Treebank (Marcus, Santorini, and Marcinkiewicz 1993). Problems with early proposals for assessing coders' agreement on discourse segmentation tasks (such as Passonneau and Litman 1993) led Carletta (1996) to suggest the adoption of the K coefficient of agreement, a variant of Cohen's κ (Cohen 1960), as this had already been used for similar purposes in content analysis for a long time.¹ Carletta's proposals were enormously influential, and K quickly became the de-facto standard for measuring agreement in Computational Linguistics not only in work on discourse (Carletta et al. 1997; Core and Allen 1997; Hearst 1997; Stolcke et al. 1997; Poesio and Vieira 1998; Di Eugenio 2000; Carlson, Marcu, and Okunowski 2003) but also for other annotation tasks (e.g., Véronis 1998; Bruce and Wiebe 1998; Stevenson and Gaizauskas 2000; Craggs and McGee Wood 2004; Nenkova and Passonneau 2004; Mieskes and Strube 2006). During this period, however, a number of questions have

¹ As we will see below, there are lots of terminological inconsistencies in the literature. Carletta uses the term κ for the coefficient of agreement, referring to Krippendorff (1988) and Siegel and Castellan (1988) for an introduction, and using Siegel and Castellan's terminology and definitions. However, Siegel and Castellan's statistic, which they call K, is actually Fleiss's generalization to more than two coders of Scott's τ , not of the original Cohen's κ ; to confuse matters further, Siegel and Castellan use the term κ to indicate the parameter which is estimated by K (i.e. a function of K with an approximately normal distribution which can be used to estimate the significance of the value of K obtained). In what follows, we will use the term κ to indicate coefficients that calculate chance agreement by looking at individual coder marginals – Cohen's original coefficient and its generalization to more than two coders – and use the term K for the coefficient discussed by Siegel and Castellan.

[Li et al, cap. 3]

- Guoliang Li, Jiannan Wang, Yudian Zheng, Ju Fan, Michael J. Franklin. *Crowdsourced Data Management – Hybrid Machine-Human Computing*, Springer, 2018

