

Lost in Replication

Un viaggio nella crisi di replicazione scientifica

Leonardo Egidì

11 Marzo 2024 - AI Talks, Hangar Teatri

Università di Trieste

I 'mille occhi': Elena Cattaneo



'Una volta ottenuti i dati finali degli esperimenti, lo scienziato non può modificarli per adeguarli alle preferenze o aspettative proprie o di una maggioranza. A volte il risultato è spiazzante, contraddice ogni previsione e ipotesi iniziale'

Elena Cattaneo, farmacologa e biologa, senatrice a vita dal 2013.



Research Article

The Fluctuating Female Vote: Politics, Religion, and the Ovulatory Cycle

Kristina M. Durante¹, Ashley Rae¹, and Vldas Griskevicius²

¹College of Business, University of Texas, San Antonio, and ²Carlson School of Management, University of Minnesota



Psychological Science
24(6) 1007–1016
© The Author(s) 2013
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0956797612466416
pss.sagepub.com
 SAGE

Da Durante, Rae, and Griskevicius (2013):

*L'ovulazione ha portato le donne **single** a diventare più liberali, meno religiose e più propense a votare per Barack Obama. Al contrario, l'ovulazione ha portato le donne **sposate** a diventare più conservatrici, più religiose e più propense a votare per Mitt Romney[...] Nel complesso, il ciclo ovulatorio non solo influenza la politica delle donne, ma sembra farlo in modo diverso per le single rispetto alle donne sposate.*

Interazione tra Fertilità (alta o bassa) e Stato relazionale (sposate e single).



3 risultati ***fondamentali*** dal paper:

- Le donne single sono più propense a votare per Barack Obama (79.3%) di quelle sposate (69.4%).

3 risultati ***fondamentali*** dal paper:

- Le donne single sono più propense a votare per Barack Obama (79.3%) di quelle sposate (69.4%). *Candidato afro-americano, prestante, intelligente... mmmmmmmhhh*

3 risultati ***fondamentali*** dal paper:

- Le donne single sono più propense a votare per Barack Obama (79.3%) di quelle sposate (69.4%). *Candidato afro-americano, prestante, intelligente... mmmmmmmhhh*
- Le donne single sono più propense a votare per Obama se sono nel periodo ad alta fertilità (86.5%) rispetto al periodo a bassa fertilità (73.7%).

3 risultati ***fondamentali*** dal paper:

- Le donne single sono più propense a votare per Barack Obama (79.3%) di quelle sposate (69.4%). *Candidato afro-americano, prestante, intelligente... mmmmmmmhhh*
- Le donne single sono più propense a votare per Obama se sono nel periodo ad alta fertilità (86.5%) rispetto al periodo a bassa fertilità (73.7%). *Eh si, ho voglia di figliare, e vorrei farlo con uno come Barack*

3 risultati ***fondamentali*** dal paper:

- Le donne single sono più propense a votare per Barack Obama (79.3%) di quelle sposate (69.4%). *Candidato afro-americano, prestante, intelligente... mmmmmmmhhh*
- Le donne single sono più propense a votare per Obama se sono nel periodo ad alta fertilità (86.5%) rispetto al periodo a bassa fertilità (73.7%). *Eh si, ho voglia di figliare, e vorrei farlo con uno come Barack*
- Le donne sposate, tuttavia, sono più propense a votare per Romney se sono in un momento ad alta fertilità (40.4%) rispetto al periodo a bassa fertilità (23.4%).

3 risultati ***fondamentali*** dal paper:

- Le donne single sono più propense a votare per Barack Obama (79.3%) di quelle sposate (69.4%). *Candidato afro-americano, prestante, intelligente... mmmmmmmhhh*
- Le donne single sono più propense a votare per Obama se sono nel periodo ad alta fertilità (86.5%) rispetto al periodo a bassa fertilità (73.7%). *Eh si, ho voglia di figliare, e vorrei farlo con uno come Barack*
- Le donne sposate, tuttavia, sono più propense a votare per Romney se sono in un momento ad alta fertilità (40.4%) rispetto al periodo a bassa fertilità (23.4%). *Mamma mia, sto 'cedro' di Mitt Romney non mi attira proprio ('azzo, e gli assomiglia pure un po' a mio marito eh...)*



Research Article



The Ancestral Logic of Politics: Upper-Body Strength Regulates Men's Assertion of Self-Interest Over Economic Redistribution

**Michael Bang Petersen¹, Daniel Sznycer², Aaron Sell^{2,3},
Leda Cosmides², and John Tooby²**

¹Department of Political Science & Government, Aarhus University; ²Center for Evolutionary Psychology, University of California, Santa Barbara; and ³School of Criminology and Criminal Justice, Griffith University

Psychological Science
24(7) 1098–1103
© The Author(s) 2013
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0956797612466415
pss.sagepub.com
 SAGE

Passa solo qualche mese dal paper precedente e ...da Petersen et al. (2013):

Abbiamo dimostrato che la forza della parte superiore del corpo negli uomini adulti moderni influenza la loro disponibilità personale alla redistribuzione del reddito e della ricchezza. Questi effetti sono stati replicati in tutte le culture e, come previsto, riscontrati solo tra i maschi.

Interazione tra status socio-economico (alto o basso) e la circonferenza delle braccia.

Risultati ***fondamentali***:

- Uomini forti con un elevato status socio-economico si oppongono alla re-distribuzione delle risorse, mentre gli uomini forti con un basso status sociale sono favorevoli alla re-distribuzione.

Braccia grosse e attitudini politiche

Risultati ***fondamentali***:

- Uomini forti con un elevato status socio-economico si oppongono alla re-distribuzione delle risorse, mentre gli uomini forti con un basso status sociale sono favorevoli alla re-distribuzione. *Quindi Braccio di Ferro, un marinaio povero in canna, vota sicuro per il M5S*



Braccia grosse e attitudini politiche

Risultati ***fondamentali***:

- Uomini forti con un elevato status socio-economico si oppongono alla re-distribuzione delle risorse, mentre gli uomini forti con un basso status sociale sono favorevoli alla re-distribuzione. *Quindi Braccio di Ferro, un marinaio povero in canna, vota sicuro per il M5S*



- La forza della parte superiore del corpo influenza il sostegno alla re-distribuzione economica nelle donne? No. *Stacce, punto.*



Research Report

Women Are More Likely to Wear Red or Pink at Peak Fertility

Alec T. Beall and Jessica L. Tracy
University of British Columbia



Psychological Science
24(9) 1837–1841
© The Author(s) 2013
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0956797613476045
pss.sagepub.com
 SAGE

Sempre 2013, stessa rivista...ci risiamo. Da **Beall and Tracy (2013)**:

[...] Le donne che sono al massimo della fertilità hanno tre volte più probabilità di indossare camicie rosse o rosa, rispetto alle donne in altri momenti del loro ciclo mestruale.

Interazione tra fertilità (alta o bassa) ed età (donne qualsiasi tra i 18 e i 47 anni e universitarie tra i 17 e i 31 anni).

Risultati ***fondamentali***:

- Le donne ad alto rischio di concepimento hanno sostanzialmente maggiori probabilità di indossare una maglietta rossa o rosa rispetto alle donne a basso rischio di concepimento.

Risultati ***fondamentali***:

- Le donne ad alto rischio di concepimento hanno sostanzialmente maggiori probabilità di indossare una maglietta rossa o rosa rispetto alle donne a basso rischio di concepimento. *Mi scusi, Lei con la maglia rossa, vorrebbe forse figliare?*

Risultati ***fondamentali***:

- Le donne ad alto rischio di concepimento hanno sostanzialmente maggiori probabilità di indossare una maglietta rossa o rosa rispetto alle donne a basso rischio di concepimento. *Mi scusi, Lei con la maglia rossa, vorrebbe forse figliare?*
- Il rischio di concepimento non ha avuto alcun effetto sulla prevalenza di qualsiasi altro colore di maglietta in entrambi i campioni. *Ah ok, niente '50 sfumature di grigio' allora, peccato.*

Early report

Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children

A J Wakefield, S H Murch, A Anthony, J Linnell, D M Casson, M Malik, M Berelowitz, A P Dhillon, M A Thomson, P Harvey, A Valentine, S E Davies, J A Walker-Smith

Summary

Background We investigated a consecutive series of children with chronic enterocolitis and regressive developmental disorder.

Methods 12 children (mean age 6 years [range 3–10], 11 boys) were referred to a paediatric gastroenterology unit with a history of normal development followed by loss of acquired skills, including language, together with diarrhoea and abdominal pain. Children underwent gastroenterological, neurological, and developmental assessment and review of developmental records. Ileocolonoscopy and biopsy sampling, magnetic-resonance imaging (MRI), electroencephalography (EEG), and lumbar puncture were done under sedation. Barium follow-through radiography was done where possible. Biochemical, haematological, and immunological profiles were examined.

Findings Onset of behavioural symptoms was associated by the parents, with measles, mumps, and rubella vaccination in eight of the 12 children, with measles infection in one child, and otitis media in another. All 12 children had intestinal abnormalities ranging from lymphoid nodular hyperplasia to pseudotuberculosis. Histology showed patchy chronic inflammation. In 11 children and reactive ileal-lymphoid hyperplasia in seven, but no granulomas. Behavioural disorders included autism (nine), disintegrative disorder (one), and possible postviral or vaccinal encephalitis (two). There were no focal neurological abnormalities and EEG and EEG tests were normal. Abnormal laboratory results were significantly raised urinary methylmalonic acid compared with age-matched controls ($p=0.03$), low haemoglobin in four children, and low serum IgA in three children.

Interpretation We identified associated gastrointestinal disease and developmental regression in a group of previously normal children, which was generally associated in time to possible environmental triggers.

Lancet 1998; 351: 637–41

See Commentary page

Inflammatory Bowel Disease Study Group, University Departments of Medicine and Histopathology (A J Wakefield, S H Murch, A Anthony, M Malik, M Berelowitz, A P Dhillon, M A Thomson, P Harvey, A Valentine, S E Davies, J A Walker-Smith)

Introduction

We saw several children who, after a period of apparent normality, lost acquired skills, including communication. They all had gastrointestinal symptoms, including abdominal pain, diarrhoea, and bloating and, in some cases, food intolerance. We describe the clinical findings, and gastrointestinal features of these children.

Patients and methods

12 children, consecutively referred to a department of paediatric gastroenterology with a history of a pervasive developmental disorder with loss of acquired skills and intestinal symptoms (abdominal pain, bloating and food intolerance), were investigated. All children were admitted to the ward for a week, accompanied by their parents.

Clinical investigations

We took history, including details of immunisations and exposure to infectious diseases, and assessed the child. In 11 cases the history was obtained by the senior clinician (JW-S). Neurological and psychiatric assessments were done by consultant staff (PH, MB) with HMS-4 criteria. Developmental assessment included a review of prospective developmental records from parents, health visitors, and general practitioners. Four children did not undergo psychiatric assessment in hospital; all had been assessed professionally elsewhere, so these assessments were used as the basis for their behavioural diagnosis.

After bowel preparation, ileocolonoscopy was performed by SHM or MAT under sedation with midazolam and pethidine. Paired frozen and formalin-fixed mucosal biopsy samples were taken from the terminal ileum; ascending, transverse, descending, and sigmoid colons, and from the rectum. The procedure was recorded by video or still images, and were compared with images of the previous seven consecutive paediatric colonoscopies (four normal colonoscopies and three on children with ulcerative colitis), in which the physician reported normal appearances in the terminal ileum. Barium follow-through radiography was possible in some cases.

Also under sedation, cerebral magnetic-resonance imaging (MRI), electroencephalography (EEG) including video, brain stem auditory, and sensory evoked potentials (where compliance made these possible), and lumbar puncture were done.

Laboratory investigations

Thyroid function, serum long-chain fatty acids, and cerebrospinal-fluid lactate were measured to exclude known causes of childhood neurodegenerative disease. Urinary methylmalonic acid was measured in random urine samples from eight of the 12 children and 14 age-matched and sex-matched normal controls, by a modification of a technique described

Vaccini e autismo: quando un articolo viene ritirato...

Da Wakefield et al. (1998):

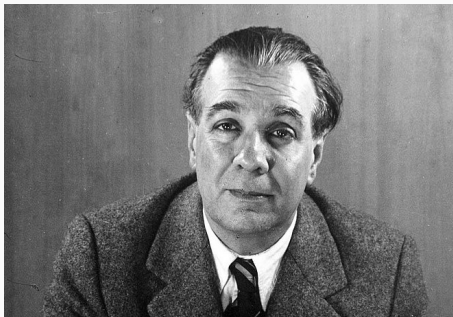
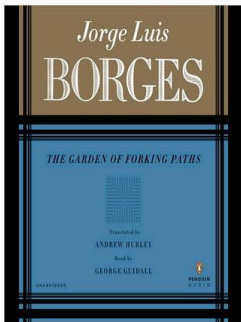
L'insorgenza di sintomi comportamentali è stata associata, da parte dei genitori, alla vaccinazione contro morbillo, parotite e rosolia in otto dei 12 bambini analizzati, con l'infezione da morbillo in un bambino e con l'otite media in un altro. Tutti e 12 i bambini presentavano anomalie intestinali, che andavano dall'iperplasia nodulare linfoide all'ulcerazione aftoide. L'istologia ha mostrato un'infiammazione cronica a chiazze nel colon in 11 bambini e un'iperplasia linfoide ileale reattiva in sette, ma nessun granuloma. I disturbi comportamentali includevano autismo (nove), psicosi disintegrativa (uno) e possibile encefalite postvirale o vaccinale (due).

'Conclusioni':

Se esiste un nesso causale tra morbillo, parotite e vaccino contro la rosolia e queste sindromi, un'incidenza in aumento potrebbe essere anticipata dopo l'introduzione di questo vaccino nel Regno Unito nel 1988.

[...] Nella maggior parte dei casi, l'insorgenza dei sintomi è avvenuta dopo la vaccinazione contro il morbillo, la parotite e la rosolia. Ulteriori indagini sono necessarie per esaminare questa sindrome e la possibile relazione con questo vaccino.

The garden of forking paths



'Pensavo a un labirinto di labirinti, a un labirinto sinuoso che abbracciava il passato e il futuro... Mi sentivo, per un periodo di tempo sconosciuto, un percettore astratto del mondo.'

Jorge Luis Borges (1941)

The garden of forking paths



John Zorn, *A Garden of Forking Paths* (2021)

The Statistical Crisis in Science

Data-dependent analysis—a “garden of forking paths”—explains why many statistically significant comparisons don’t hold up.

Andrew Gelman and Eric Loken

There is a growing realization that reported “statistically significant” claims in scientific publications are routinely mistaken. Researchers typically express the confidence in their data in terms of p -value: the probability that a perceived result is actually the result of random variation. The value of p (for “probability”) is a way of measuring the extent to which a data set provides evidence against a so-called null hypothesis. By convention, a p -value below 0.05 is considered a meaningful refutation of the null hypothesis; however, such conclusions are less solid than they appear.

The idea is that when p is less than some prespecified value such as 0.05, the null hypothesis is rejected by the data, allowing researchers to claim strong evidence in favor of the alternative. The concept of evidence was solid

a short mathematics test when it is expressed in two different contexts, involving either healthcare or the military. The question may be framed nonspecifically as an investigation of possible associations between party affiliation and mathematical reasoning across contexts. The null hypothesis is that the political context is irrelevant to the task, and the alternative hypothesis is that context matters and the difference in performance between the two parties would be different in the military and healthcare contexts.

At this point a huge number of possible comparisons could be performed, all consistent with the researcher’s theory. For example, the null hypothesis could be rejected (with statistical significance) among men and not among women—explicable under the theory that men are more ideological than women. The pattern could be found

This *multiple comparisons* issue is well known in statistics and has been called “ p -hacking” in an influential 2011 paper by the psychology researchers Joseph Simmons, Leif Nelson, and Uri Simonsohn. Our main point in the present article is that it is possible to have multiple potential comparisons (that is, a data analysis whose details are highly contingent on data, invalidating published p -values) without the researcher performing any conscious procedure of fishing through the data or explicitly examining multiple comparisons.

How to Test a Hypothesis

In general, we could think of four classes of procedures for hypothesis testing: (1) a simple classical test based on a unique test statistic, T , which when applied to the observed data yields T_{obs} , where α represents the

Andrew Gelman, Erik Loken (2014)

Uno statistico? Sa bene solo una cosa



'Come statistico non sono esperto di X , per qualsiasi X che mi capita di studiare. Credo che uno statistico possa essere veramente esperto e saperne qualcosa solo di un ambito tra i vari che studia.'

Andrew Gelman, statistico

I cavalieri dell'apocalisse



'Ci sono quattro 'cavalieri' che remano contro alla replicazione scientifica: bias da pubblicazione, p-hacking, campioni piccoli e story-teling post sperimentale.'

Dorothy Bishop, psicologa

The NEW ENGLAND JOURNAL *of* MEDICINE

ESTABLISHED IN 1812

DECEMBER 31, 2020

VOL. 383 NO. 27

Safety and Efficacy of the BNT162b2 mRNA Covid-19 Vaccine

Fernando P. Polack, M.D., Stephen J. Thomas, M.D., Nicholas Kitchin, M.D., Judith Absalon, M.D.,
Alejandra Gurtman, M.D., Stephen Lockhart, D.M., John L. Perez, M.D., Gonzalo Pérez Marc, M.D.,
Edson D. Moreira, M.D., Cristiano Zerbini, M.D., Ruth Bailey, B.Sc., Kena A. Swanson, Ph.D.,
Satrajit Roychoudhury, Ph.D., Kenneth Koury, Ph.D., Ping Li, Ph.D., Warren V. Kalina, Ph.D., David Cooper, Ph.D.,
Robert W. Frenck, Jr., M.D., Laura L. Hammitt, M.D., Özlem Türeci, M.D., Haylene Nell, M.D., Axel Schaefer, M.D.,
Serhat Ünal, M.D., Dina B. Tresnan, D.V.M., Ph.D., Susan Mather, M.D., Philip R. Dormitzer, M.D., Ph.D.,
Uğur Şahin, M.D., Kathrin U. Jansen, Ph.D., and William C. Gruber, M.D., for the C4591001 Clinical Trial Group*

ABSTRACT

Un buon esempio di replicazione scientifica: Pfizer

Table 2. Vaccine Efficacy against Covid-19 at Least 7 days after the Second Dose.*

Efficacy End Point	BNT162b2		Placebo		Vaccine Efficacy, % (95% Credible Interval)‡	Posterior Probability (Vaccine Efficacy >30%)§
	No. of Cases	Surveillance Time (n)†	No. of Cases	Surveillance Time (n)†		
Covid-19 occurrence at least 7 days after the second dose in participants without evidence of infection	(N=18,198)		(N=18,325)		95.0 (90.3–97.6)	>0.9999
	8	2.214 (17,411)	162	2.222 (17,511)		
Covid-19 occurrence at least 7 days after the second dose in participants with and those without evidence of infection	(N=19,965)		(N=20,172)		94.6 (89.9–97.3)	>0.9999
	9	2.332 (18,559)	169	2.345 (18,708)		

* The total population without baseline infection was 36,523; total population including those with and those without prior evidence of infection was 40,137.

† The surveillance time is the total time in 1000 person-years for the given end point across all participants within each group at risk for the end point. The time period for Covid-19 case accrual is from 7 days after the second dose to the end of the surveillance period.

‡ The credible interval for vaccine efficacy was calculated with the use of a beta-binomial model with prior beta (0.700102, 1) adjusted for the surveillance time.

§ Posterior probability was calculated with the use of a beta-binomial model with prior beta (0.700102, 1) adjusted for the surveillance time.

	CASES
Vaccinated	$n_v = 8$
Controls	$n - n_v = 162$
	$n = 170$

$$\theta \sim \text{Beta}(0.700102, 1)$$



$$\theta | \text{Data} \sim \text{Beta}(0.700102 + 8, 1 + 162)$$

Assessing replication success via skeptical mixture priors

Guido Consonni* and Leonardo Egidi†

December 30, 2023

Abstract

There is a growing interest in the analysis of replication studies of original findings across many disciplines. When testing a hypothesis for an effect size, two Bayesian approaches stand out for their principled use of the Bayes factor (BF), namely the replication BF and the skeptical BF. In particular, the latter BF is based on the skeptical prior, which represents the opinion of an investigator who is unconvinced by the original findings and wants to challenge them. We embrace the skeptical perspective, and elaborate a novel mixture prior which incorporates skepticism while at the same time controlling for prior-data conflict within the original data. Consistency properties of the resulting skeptical mixture BF are provided together with an extensive analysis of the main features of our proposal. Finally, we apply our methodology to data from the Social Sciences Replication Project. In particular we show that, for some case studies where prior-data conflict is an issue, our method uses a more realistic prior and leads to evidence-classification for replication success which differs from the standard skeptical approach.

Keywords: Bayes factor, Bayesian hypothesis testing, consistency, prior-data conflict, replication studies, reverse-Bayes

Consonni and Egidi (2023), under review.

‘Pensando di poter sempre estrarre il segnale dal rumore, la statistica “vince” quando scopre effetti statisticamente significativi, quando trova un p -value minore di 0.05. I veri progressi richiederanno che i ricercatori facciano previsioni più capaci di dimostrare le loro teorie e investano in misurazioni più precise che presentino, in molti casi, confronti individuali. Un passo cruciale è andare oltre l'alchimia delle affermazioni binarie su “un effetto” o “nessun effetto” con solo un valore P a dividerli. Accettare l'incertezza, abbracciare la variabilità’.

Sole-lease Via Angelo Trossi 8, Milano

EDIZIONI TELETRASMESSE: BCS Produzioni Milano S.p.A. 20060 Pessano con Bornago - Via R. Luxemburg - Tel. 02-6281.8298 • BCS Produzioni S.p.A. 00195 Roma - Via Ciamatti 102/103 - Tel. 06-68.85.8427 • BCS Produzioni Padova S.p.A. 37100 Padova - Corso Sta-

e 0,79; ven. Corvera + 7 + CoV'en. e 0,80 = e 0,50 + e 0,79; sub. Corvera + la Donna + CoV'en. e 0,80 + e 0,50 + e 0,79; dom. Corvera + la Donna + CoV'en. e 0,80 + e 0,50 + e 0,79. In Trentino Alto Adige, non sono stati censiti né le Corvera, né CoV'en, né CoV'om, né CoV'ib, né So, né po.

«Capire il lavoro di ricerca che c'è dietro la didattica»

In questo periodo, l'unica proiezione visibile del mondo dell'istruzione per media e governanti avviene sul piano della didattica. Come se l'unica funzione del docente/ricercatore universitario fosse quella di insegnare qualche ora, poi per tornare nei suoi agi di staltale a bassa produttività; otto mesi di pandemia senza quasi mai sentir nominare la ricerca, di fatto l'unico discrimine tra buona e cattiva università. Insegnare ai tempi del Covid è duro, tra monitoraggio di telecamere, schermi, microfoni e connessioni: ma poco e in confronto alla frustrazione di sentirsi esclusivamente etichettati come «vettori di contenuti». Dietro questa veicolazione ci sono ricerca, studio, abnegazione; su un piano ancora più invisibile e posteriore, supervisione di studenti, riunioni, verbali, convegni, seminari, e

molto altro. In questi mesi, questo polimorfismo sta venendo sempre più schiacciato e proiettato in un'unica dimensione, come avviene nel romanzo *Flatlandia*. La pandemia blocca i contatti ma distanza forse anche la percezione dei ruoli, degli stessi quotidiani profusi per rendersi credibili davanti a una platea di ragazzi. Senza ricerca non saremmo buoni docenti, ma pareremmo un po' come il protagonista de *Le nuvole* di Aristofane, Strepsade, indottrinato alla bell'e meglio da Socrate e incapace di distinguere tra il vero e il falso, di scegliere da lui quanto gli spetta. Ecco, i ragazzi a cui insegniamo sono creditori di una fatica traslucida, di buona ricerca.

Leonardo Egidi, Ricercatore universitario e docente in Statistica, Università di Trieste



Il lettore, docente universitario, precisa che i prof non sono solo vettori di contenuti, ma il loro lavoro è anche di ricerca e molto altro ancora.

‘L’unico modo per imparare davvero una materia è insegnarla’

25