# An overview of the elementary statistics of correlation, *R*-Squared, cosine, sine, Xur, Yur, and regression through the origin, with application to votes and seats for parliament

Thomas Colignatus

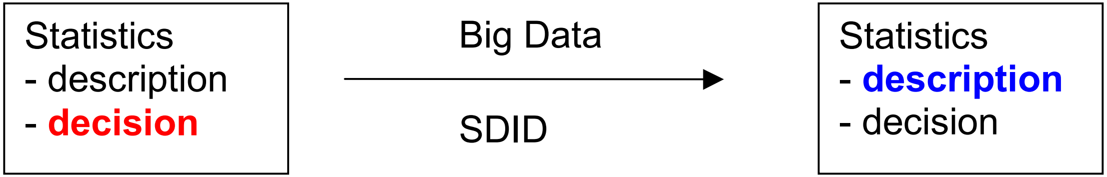
*Sheets for the Politicologenetmaal*, Leiden, June 8 2018

Paper at:

https://mpra.ub.uni-muenchen.de/86307/

https://zenodo.org/record/1228640

Statistics:
From a focus on ***decision*** (hypothesis testing)
to a focus on ***description***



Statistics
- description
- **decision**

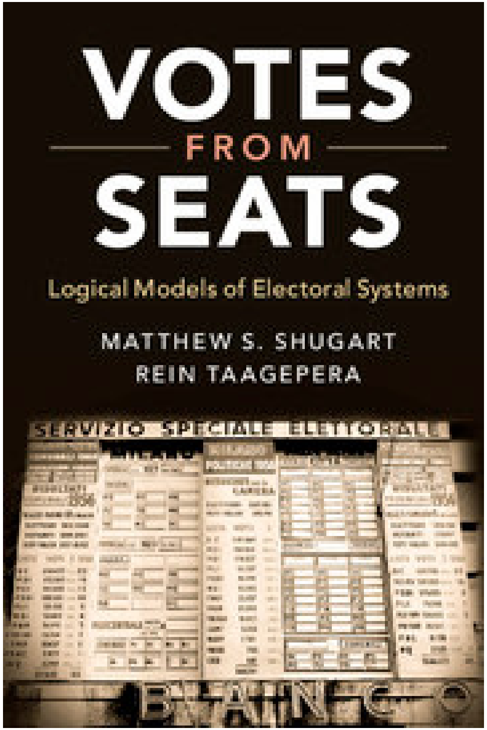
Big Data

SDID

Statistics
- **description**
- decision

SDID = Sine-Diagonal Inequality or Disproportionality

A measure of distance between votes and seats


VOTES
FROM
SEATS
Logical Models of Electoral Systems
MATTHEW S. SHUGART
REIN TAAGEPERA
SERVIZIO SPECIALE ELETTORALE

2017

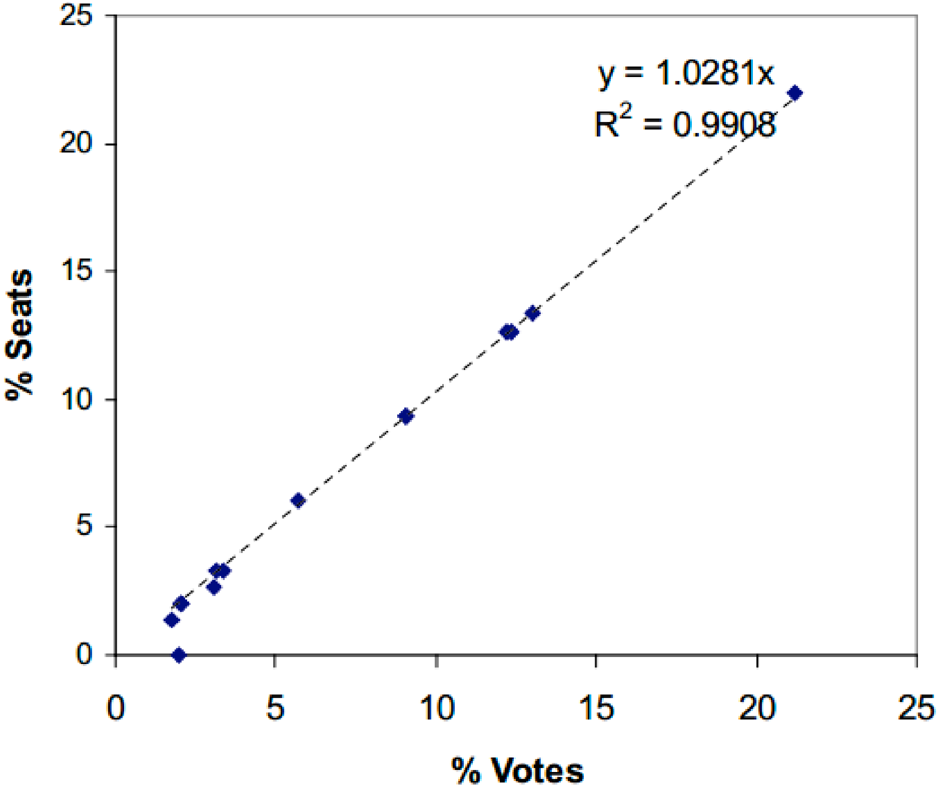
“Thus the book is a rare *scientific* book about politics, and should set a methodological standard for all social sciences.” (p320)

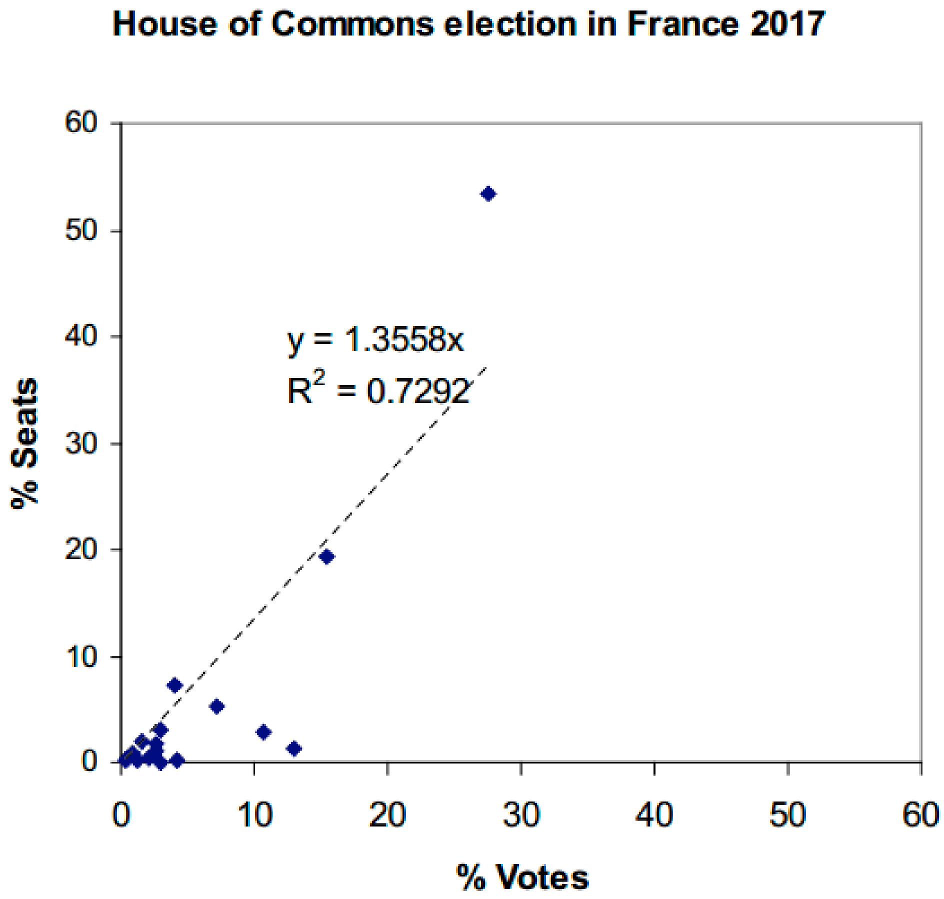
On statistical significance:

“This can produce valuable insights, but these so-called “empirical models” are not really models at all. (...) Every peasant in Galileo’s time knew the direction in which things fall - but Galileo felt the need to predict more than direction.” (p324).

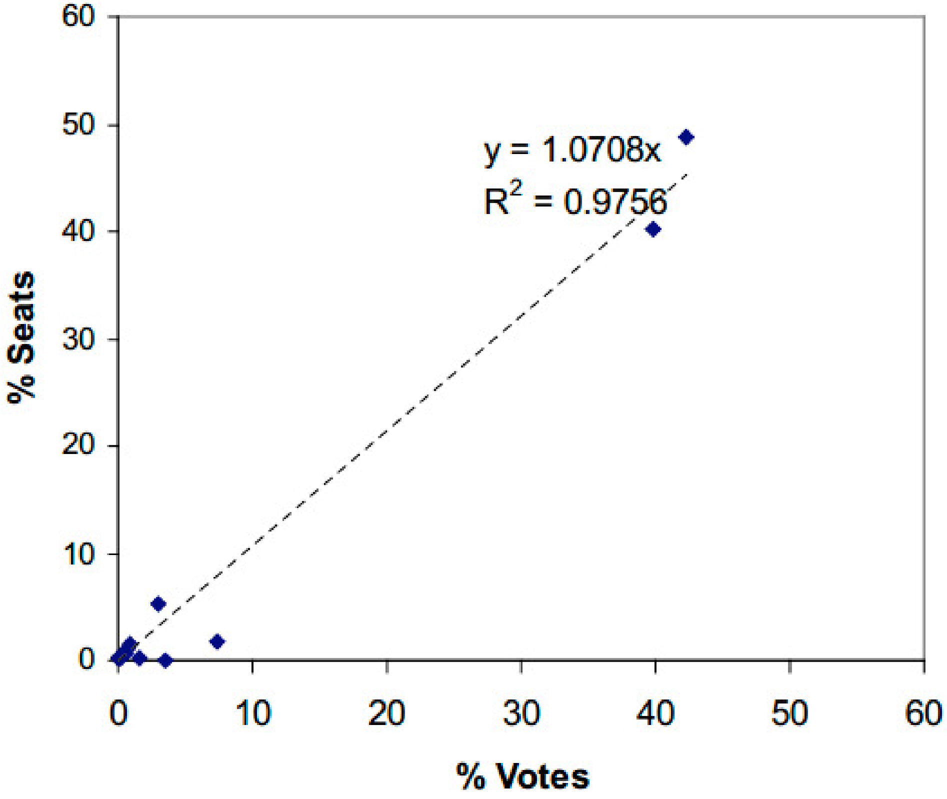
Better look at the effect size.

**House of Commons election in Holland 2017**




House of Commons election in France 2017
% Seats
60
50
40
30
20
10
0
% Votes
0
10
20
30
40
50
60
y = 1.3558x
$R^2$ = 0.7292

House of Commons election in UK 2017



# New approach

For votes and seats:

- for standardised variables the regression coefficient *b* is also the correlation *R:* thus *R* is more fundamental

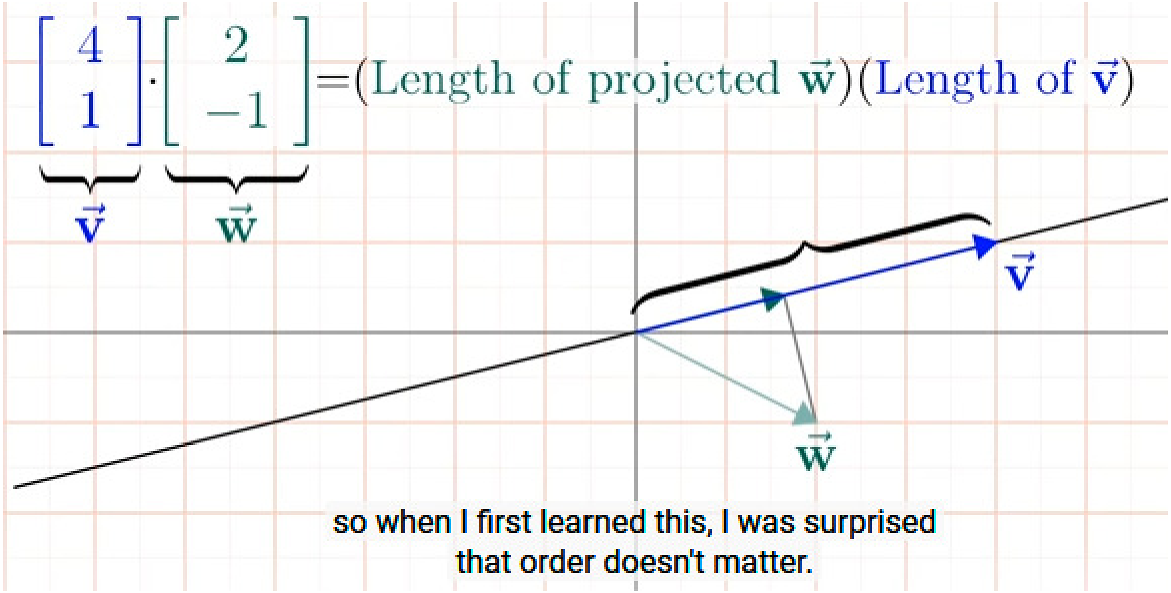
- use $d = \sqrt{(1 - R^2)}$ as a distance measure

- enhance sensitivity by using $\sqrt{d}$ (take the sqrt a second time)

- resolve issues on District Representation (DR) and Equal / Proportional Representation (EPR), both on content and measurement

# √*R*-squared = *R* = Correlation = Cosine = Projection onto *x*-axis

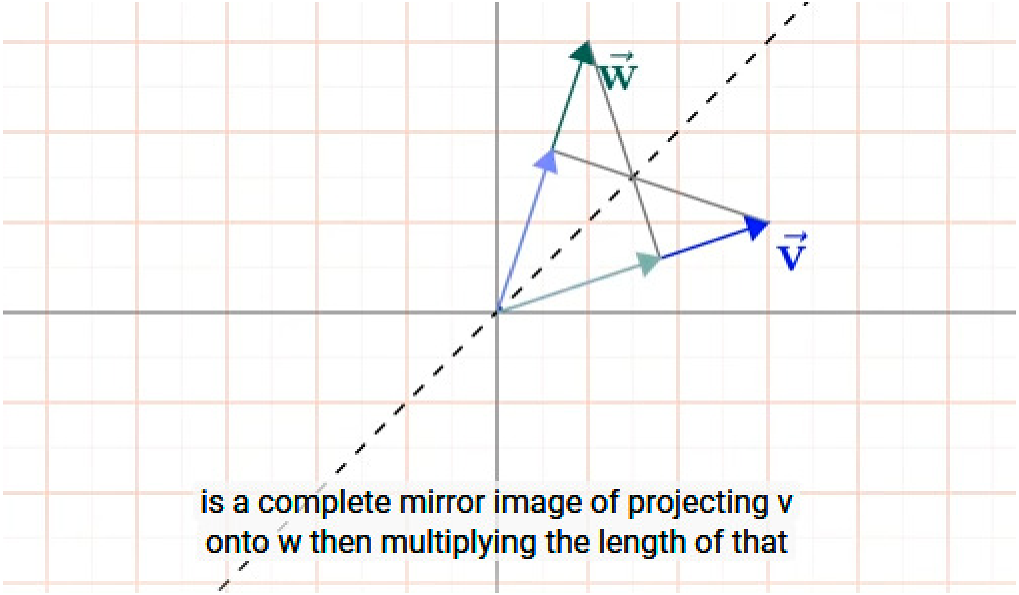
3Blue1Brown: https://www.youtube.com/watch?v=LyGKycYT2v0&feature=youtu.be&t=2m10s



$$\underbrace{\begin{bmatrix}4\\1\end{bmatrix}}_{\vec{\mathbf{v}}}\cdot\underbrace{\begin{bmatrix}2\\-1\end{bmatrix}}_{\vec{\mathbf{w}}}=(\text{Length of projected }\vec{\mathbf{w}})(\text{Length of }\vec{\mathbf{v}})$$

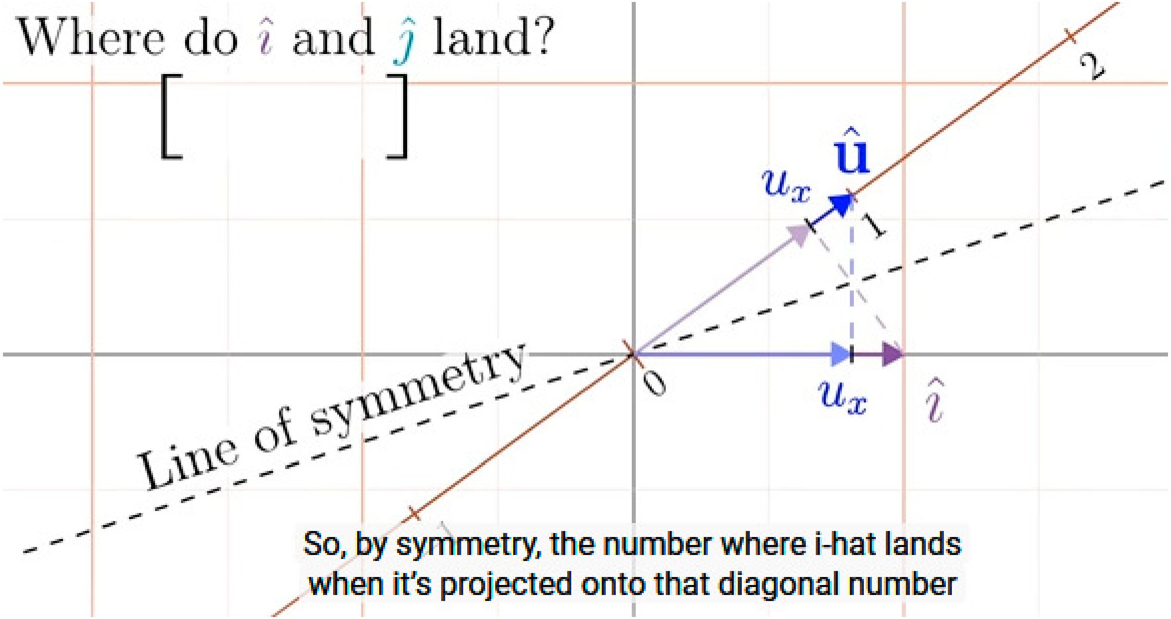
so when I first learned this, I was surprised
that order doesn't matter.

# Symmetry for two vectors of equal length



is a complete mirror image of projecting v onto w then multiplying the length of that

Normalise, take one vector as the *x*-axis, and project the other onto this



So, by symmetry, the number where i-hat lands when it's projected onto that diagonal number

LHS is normalised to 1: the cosine is also the regression coefficient *b*

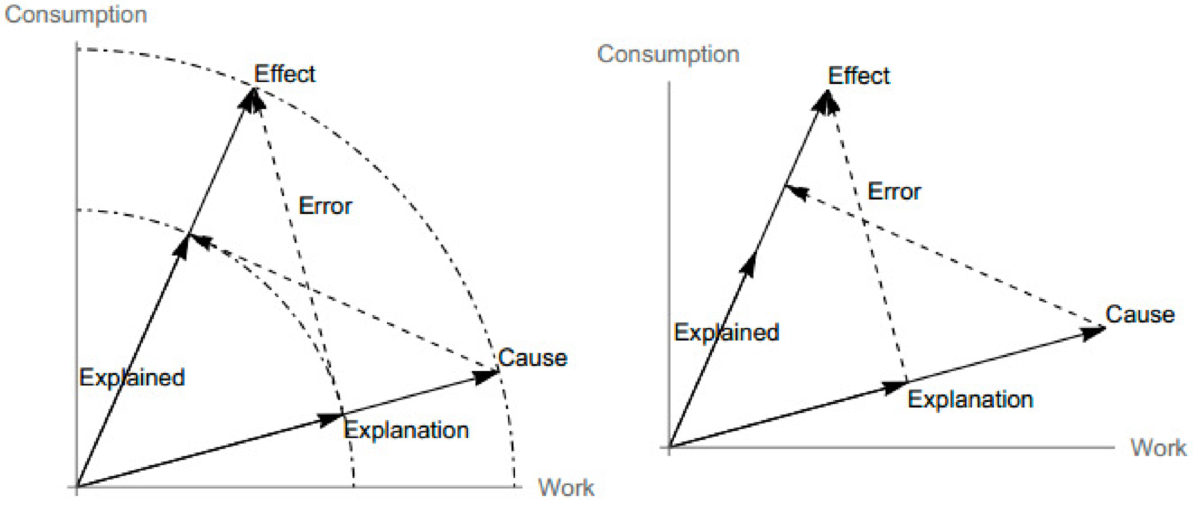


Figure 2: Projection of Effect {4, 9} on Cause {11, 3}[16]

## Lessons for *statistics* and *education of statistics*:

- Measure of association: Cosine = correlation = √*R*-squared (known)
- Measure of distance: Sine = √(1 – Cos$^2$) (not used now)
- Regression: minimisation of the sum of squared errors (known), and normal distribution of errors for hypothesis testing (known)
- Regression: look at the angle between the vectors, with the cosine as the projection (known), and describe the disproportionality (new)
- Regression: not only *hypothesis testing* but also *description*
- Help political science become a *science* on *votes and seats*

Three old measures of the distance between votes and seats. The new SDID. Shares normalised to 10. *S* = number of seats.

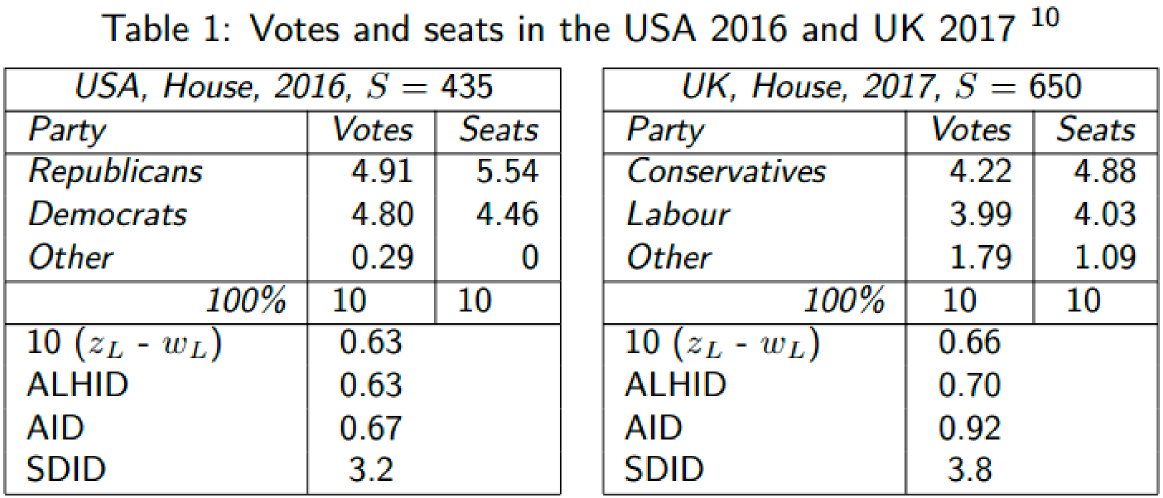


Table 1: Votes and seats in the USA 2016 and UK 2017 [10]

| *USA, House, 2016, $S = 435$* | | |
|---|---|---|
| *Party* | *Votes* | *Seats* |
| *Republicans* | 4.91 | 5.54 |
| *Democrats* | 4.80 | 4.46 |
| *Other* | 0.29 | 0 |
| *100%* | 10 | 10 |
| 10 ($z_L$ - $w_L$) | 0.63 | |
| ALHID | 0.63 | |
| AID | 0.67 | |
| SDID | 3.2 | |

| *UK, House, 2017, $S = 650$* | | |
|---|---|---|
| *Party* | *Votes* | *Seats* |
| *Conservatives* | 4.22 | 4.88 |
| *Labour* | 3.99 | 4.03 |
| *Other* | 1.79 | 1.09 |
| *100%* | 10 | 10 |
| 10 ($z_L$ - $w_L$) | 0.66 | |
| ALHID | 0.70 | |
| AID | 0.92 | |
| SDID | 3.8 | |

# Traditional distances for votes and seats

- Absolute difference / Loosemore-Hanby (ALHID): 10 Sum[Abs[$z$ - $w$] / 2]. The division by 2 corrects for double counting. An outcome of 1 means that one seat in a House of 10 seats is relocated from equality / proportionality.

- Euclid / Gallagher (EGID): 10 $\sqrt{\text{Sum}\left[(z-w)^2/\,2\right]} = 10\ ||z-w||/\sqrt{2}$, with the first form for comparisons. For two parties this equals ALHID.

- $\chi^2$ / Webster / Sainte-Laguë (CWSID): 10 Sum$[w\,(z/w-1)^2]$ = 10 Sum$[(z-w)^2/w]$. The Chi-Square expression has nonzero $w$. One can compare CWSID with ALHID = 10 Sum[$w$ Abs[$z/w$ - 1] / 2] and EGID = 10 $\sqrt{\text{Sum}\left[w^2(z/w-1)^2/\,2\right]}$

- The difference in shares for the "largest" party, i.e. with the most seats: 10 ($z_L$ - $w_L$). This is an easy, rough and ready indicator with some history in the literature, and Shugart and Taagepera (2017) p143 show remarkably that EGID ≈ 10 ($z_L$ - $w_L$).

ALHID: blue. AID = 10 $\theta$ / 90º: yellow. Sine: green. SDID = sgn √Sine: red

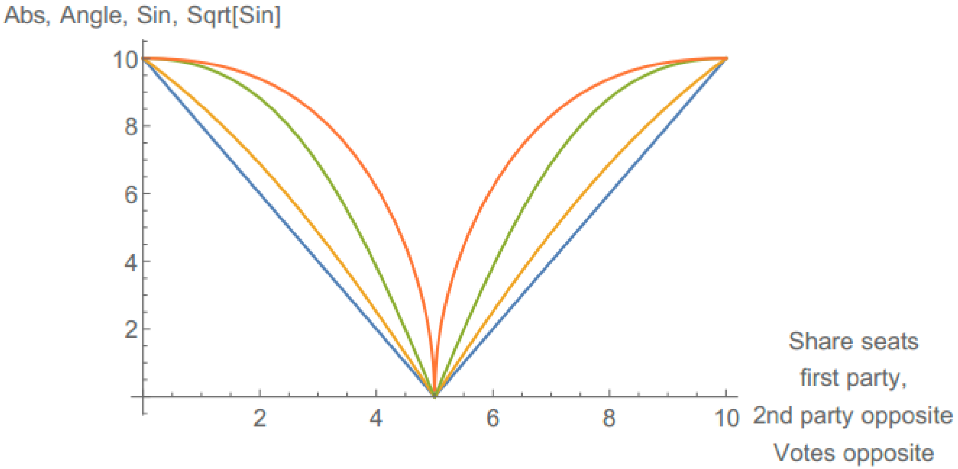
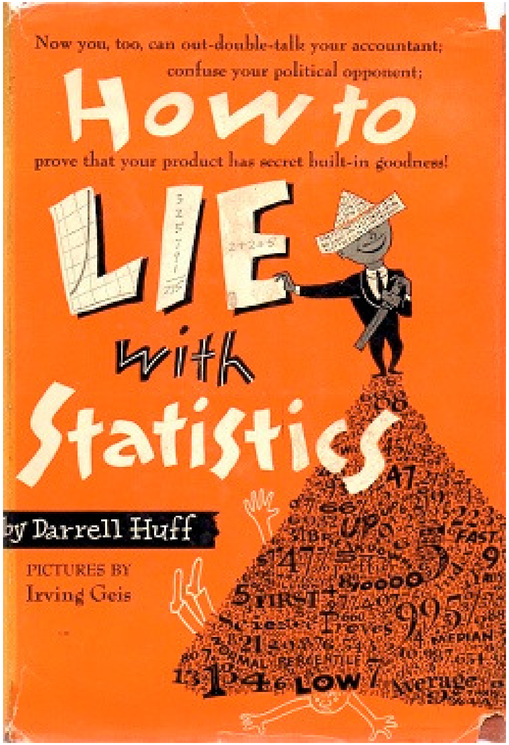


Figure 1: Plot of $d[votes, seats]$ for $votes = 10$ - $seats$ and $seats = \{t, 10 - t\}$, for $d =$ Abs/2, AngularID, Sine, and |SDID| (eliminating the latter's negative sign)

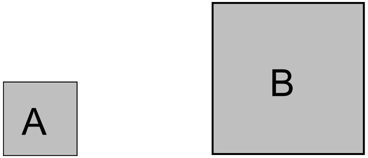

Now you, too, can out-double-talk your accountant;
confuse your political opponent;
How to
prove that your product has secret built-in goodness!
LIE
with
Statistics
by Darrell Huff
PICTURES BY
Irving Geis

Statistical integrity

- Do not distort
- Richter scale: log

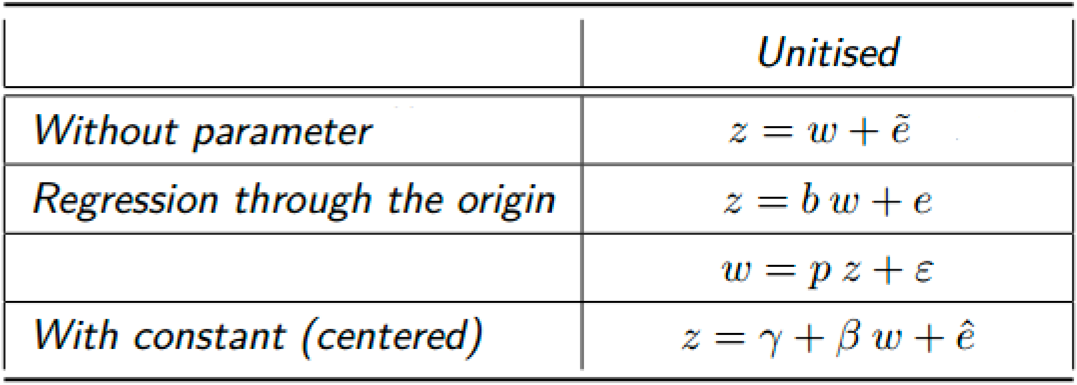
Key: *information*

B is twice as high as A



# Three models in descriptive statistics (*with errors* !)

- *w* = *v* / *V* = share of votes
- *z* = *s* / *S* = share of seats



| | *Unitised* |
|---|---|
| *Without parameter* | $z = w + \tilde{e}$ |
| *Regression through the origin* | $z = b\,w + e$ |
| | $w = p\,z + \varepsilon$ |
| *With constant (centered)* | $z = \gamma + \beta\,w + \hat{e}$ |

Regression through the origin (RTO) is better than constant (Pearson). Traditional ALHID & EGID have no parameter. SDID has *b* and *p.*

# (Likely) New finding

- Traditional ALHID & EGID divide by 2 to remain in [0, 1]
- Sine & SDID don’t need such adjustment

Derived is this relationship:

$$\text{Sin}^2 = (\tilde{e}'\tilde{e}/z'z - h)/(1 - h) \qquad h = (1 - b)^2 \;/\; b^2 = (1/b - 1)^2$$

$$\text{Sin}^2 = (\tilde{e}'\tilde{e} \;/\; w'w - g) \;/\; (1 - g) \qquad g = (1/p - 1)^2$$

$$\text{EGID} = \sqrt{\tilde{e}'\tilde{e}/2}$$

The latter was the heuristic that started the Colignatus (2018b) paper. Taking the geometric average $\sqrt{b\,p}$ gave the recognition that this gave the same mathematical expression of the cosine as well, or Cos$[v, s] = \sqrt{b\,p}$. At some point it appeared that the role of the cosine was more important by itself, and thus not regarded as a slope, as it generates the inequality / disproportionality measure Sin $= \sqrt{1 - b\,p}$. This again was first seen as a *slope-diagonal* deviation measure but eventually the name *sine-diagonal* is more accurate. This double nature of cosine and sine may be illustrated by Rubin's Vase, see **Figure** 3.

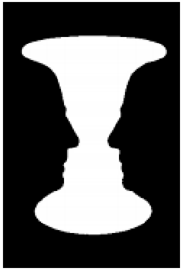


Figure 3: Rubin's Vase [25]

# District Representation vs Equal / Proportional Representation

- In EPR, “representation” for Parliament means “standing for the people who have voted for you, by marking your name or party”. In EPR, a candidate gets a seat when the natural quota $Q = V/S$, the national average votes per seat, is covered, while this criterion is only lowered for the remainder seats. Thus there tends to be full backing by those like-minded.

- In DR, there is (a) the confusion in SSD between the single seat election and the multiple seats election, or (b) the confusion also in larger district magnitudes $(1 < M \ll S)$ between a proper election and a contest. The *de iure* House of Representatives is *de facto* a House of District Winners, and *de facto* not a House of Representatives in the sense of EPR. DR “assumes” (doesn't think through) the green cheese Moon that the winner of a district seat “represents” all conflicting interests of both who voted for him or her and who explicitly didn't.

Scientific integrity: Are these elections or contests ?

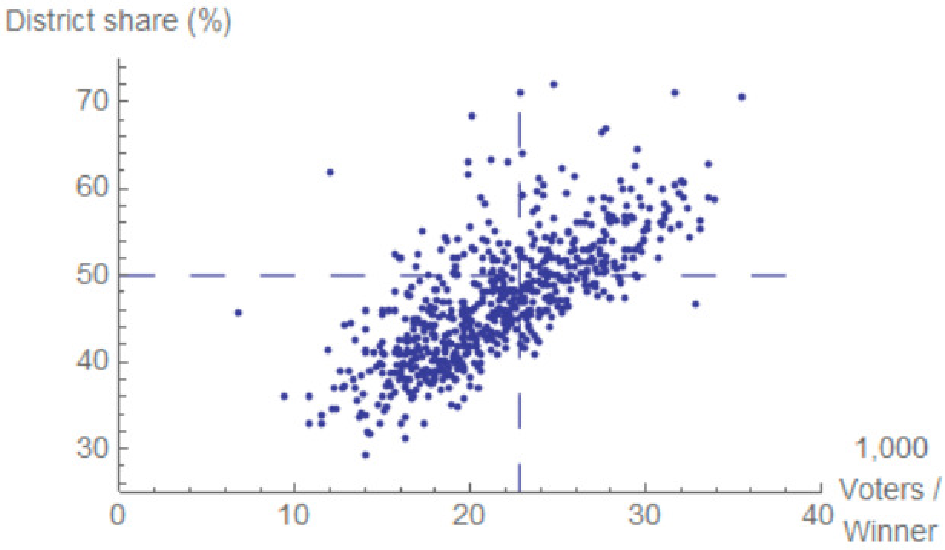
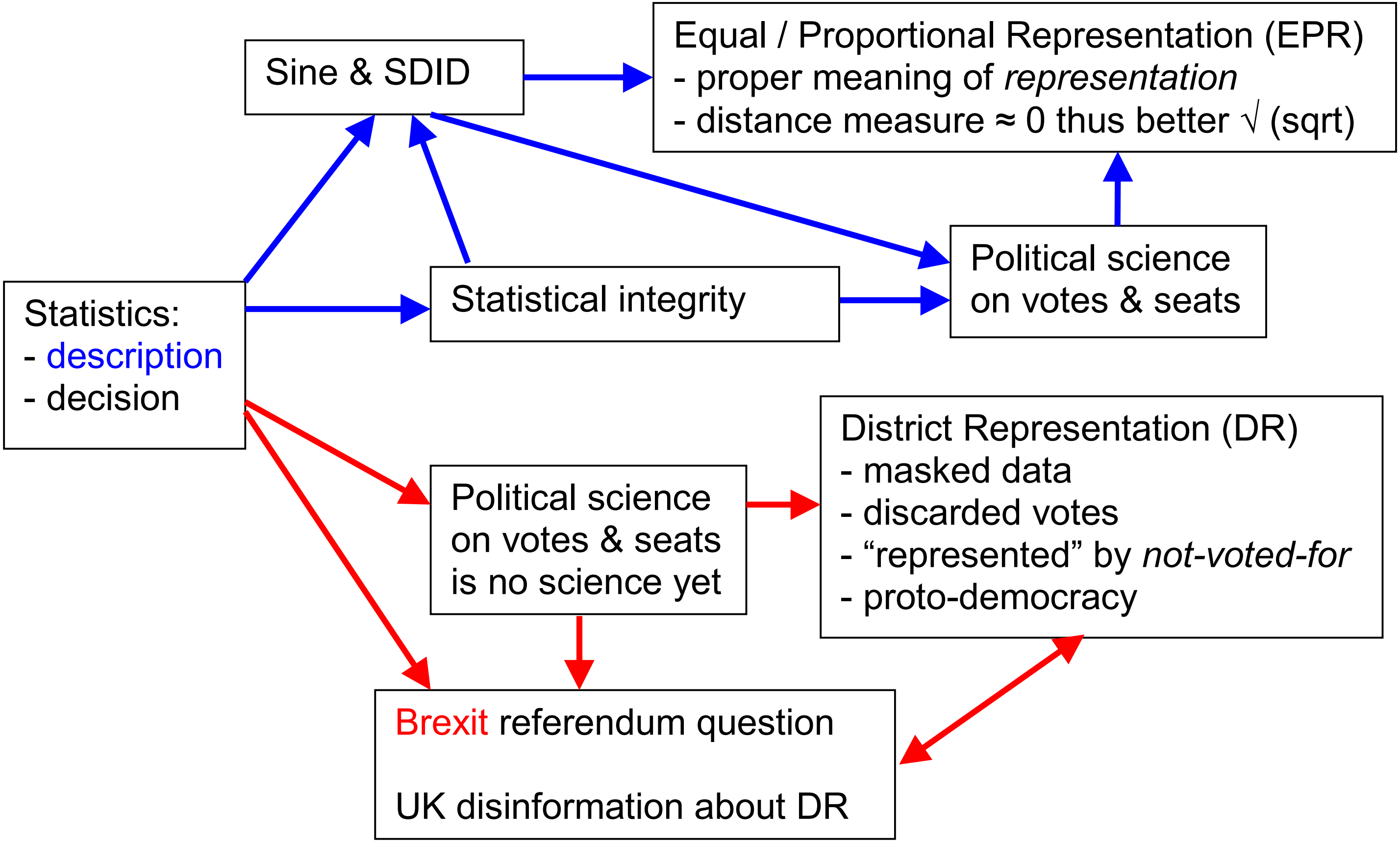


Figure 4: MPs of UK 2010: Winning % (District share) per votes *won* per seat



Sine & SDID

Equal / Proportional Representation (EPR)
- proper meaning of *representation*
- distance measure ≈ 0 thus better √ (sqrt)

Statistics:
- description
- decision

Statistical integrity

Political science on votes & seats

Political science on votes & seats is no science yet

District Representation (DR)
- masked data
- discarded votes
- “represented” by *not-voted-for*
- proto-democracy

Brexit referendum question

UK disinformation about DR

# Conclusions

(1) SDID can be used as a measure for votes and seats, comparable to the Richter scale for earthquakes

(2) Statistics education better inserts a focus on the angle and trigonometry and Regression through the Origin (RTO), with descriptive statistics, before looking at the sum of squares with the framework of hypothesis testing. This fits the shift in Big Data

(3) Statistics ethics comes to the fore on the issue of providing proper information: (a) the additional Sqrt in SDID, like the Log in the Richter scale, (b) the difference between DR and EPR must be described as what it is: the difference beween contests and elections

(4) Shugart & Taagepera (2017) make a huge step forward but still suffer from the confusion about DR and EPR, and still isn’t science