dati

dati

informazioni raccolte in modo sistematico

in un contesto e per uno scopo

dati non è il plurale di aneddoto

osservazioni anedottiche

25

Nicola

41

blue

Gerard

185

30

variabili

etichette per insiemi di dati

in studi sperimentali o osservativi

dipendenti o indipendenti

variabili

possono essere numeri o categorie, between o within

all’interno di queste distinzioni fondamentali ci sono tantissime sfumature

formato rettangolare standard

studentiTDA.xls

	A	B	C	D	E	F	G	H	I
1	nome	cognome	n	eta	sesso	ncorsipn	altezza(cm)	LT	altre lauree
2	asya	barbieri	1	23	f	0	160	parma	no
3	mauro	belluardo	2	24	m	1	185	roma	no
4	francesco	bellucci	3	23	m	0	182	parma	no
5	francesca	bertolotti	4	24	f	1	159	parma	no
6	jennika	caffelli	5	26	f	2	165	parma	no
7	stefano	colia	6	30	m	2	186	roma	logopedia
8	giorgia	colopi	7	23	f	0	162	parma	no
9	valentina	dallari	8	23	f	0	165	parma	no
10	cristian	datteri	9	23	m	0	165	parma	no
11	danila	de matteis	10	24	f	1	166	roma	no
12	chiara	dell'amico	11	22	f	2	160	massa	no
13	raffaele	di lorenzo	12	38	m	2	173	parma	sociologia
14	fabio	fogoun	13	24	m	1	175	parma	no
15	roberta	galentino	14	23	f	0	163	parma	no
16	marta	ghiglione	15	22	f	0	170	parma	no
17	daniela	licata	16	23	f	0	160	parma	no
18	silvia	manini	17	25	f	1	180	parma	no
19	alessandro	marrazzo	18	25	m	2	179	roma	no
20	margherita	martani	19	43	f	3	160	parma	economia
21	luca	marzari	20	29	m	1	183	parma	no
22	erica	morciano	21	23	f	0	170	parma	no
23	ilenia	ognibene	22	23	f	0	170	parma	no
24	serena	orrù	23	26	f	2	160	parma	no
25	emanuela	pedrinelli	24	22	f	0	160	parma	no
26	francesca	siri	25	23	f	0	160	parma	no
27	selena	tanzi	26	24	f	2	163	parma	no
28									
29									

R

creare un dataframe

esempio di uso di una funzione: pie()

decifrare help(pie):

descrizione, uso, argomenti, vedi anche, esempi e note

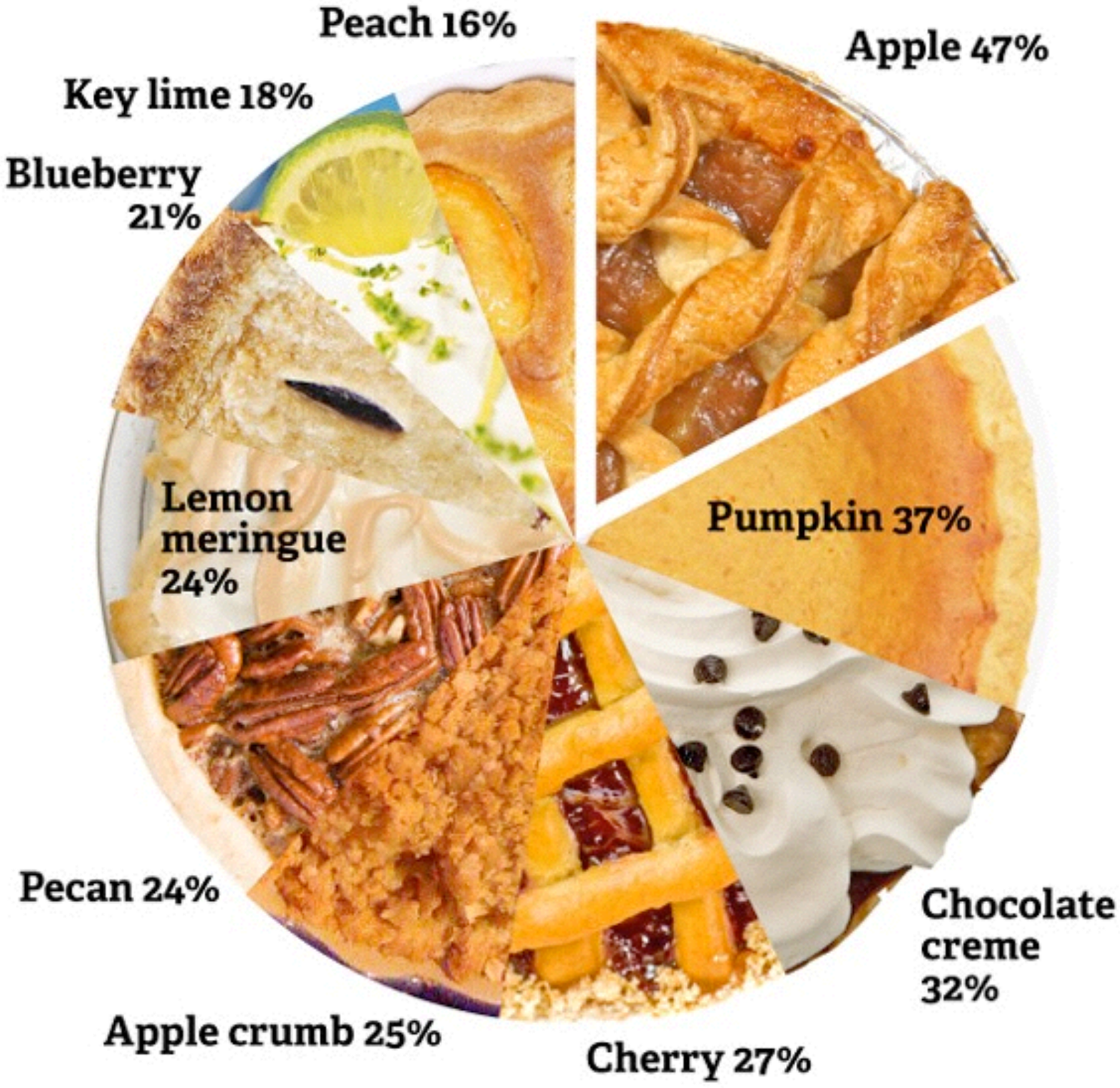
la nota su help(pie)

Note

Pie charts are a very bad way of displaying information. The eye is good at judging linear measures and bad at judging relative areas. A bar chart or dot chart is a preferable way of displaying this type of data.

Cleveland (1985), page 264: “Data that can be shown by pie charts always can be shown by a dot chart. This means that judgements of position along a common scale can be made instead of the less accurate angle judgements.” This statement is based on the empirical investigations of Cleveland and McGill as well as investigations by perceptual psychologists.


Pie I have eaten
Pie I have not yet eaten



http://www.livescience.com/33111-favorite-pie-america.html

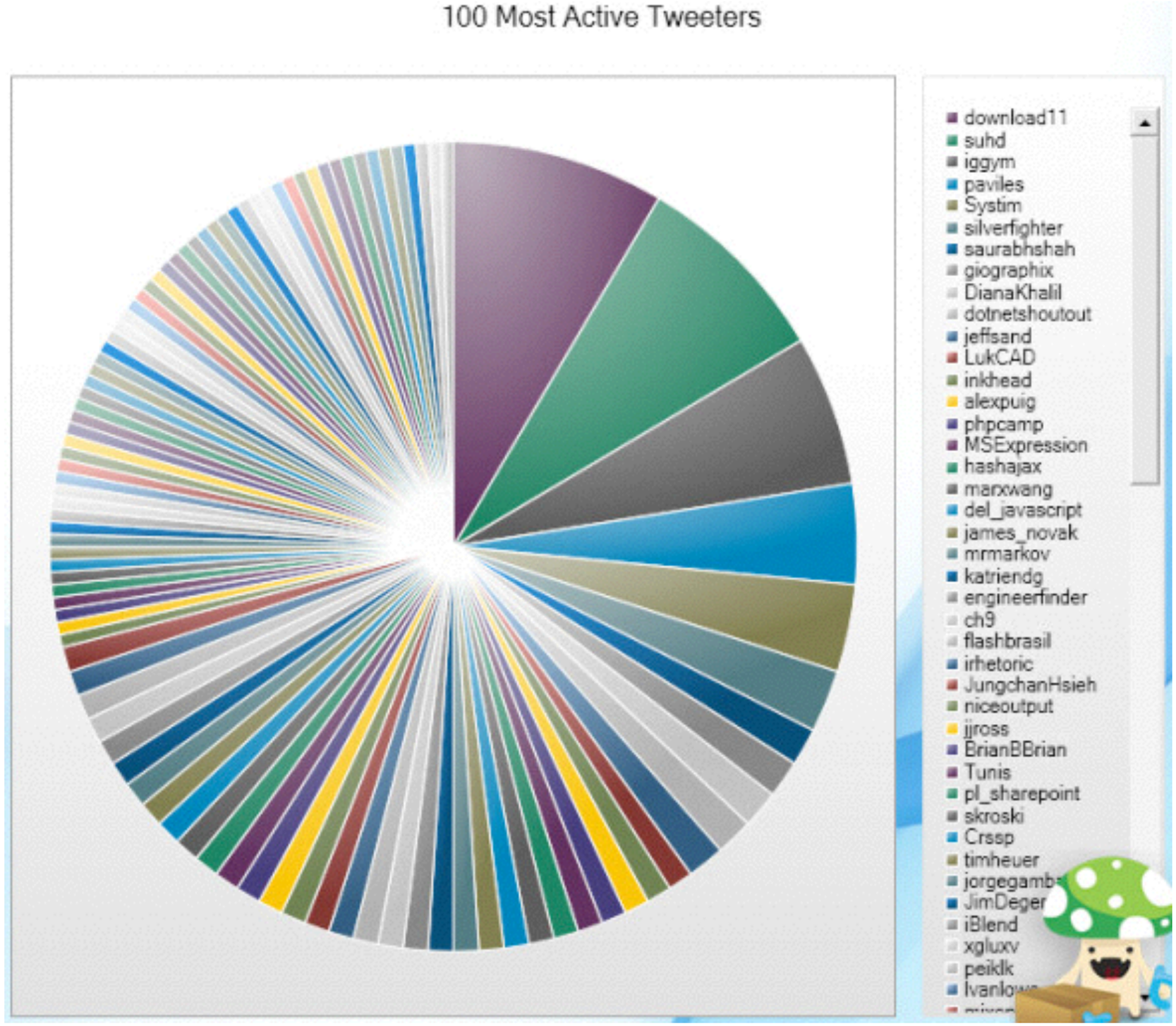
il venerdì
numero 1041
29 febbraio 2008
di Repubblica

30%
GLI INTERVISTATI
CHE NON RITENGONO
L'UNITÀ NAZIONALE
UN BENE IRRINUNCIABILE

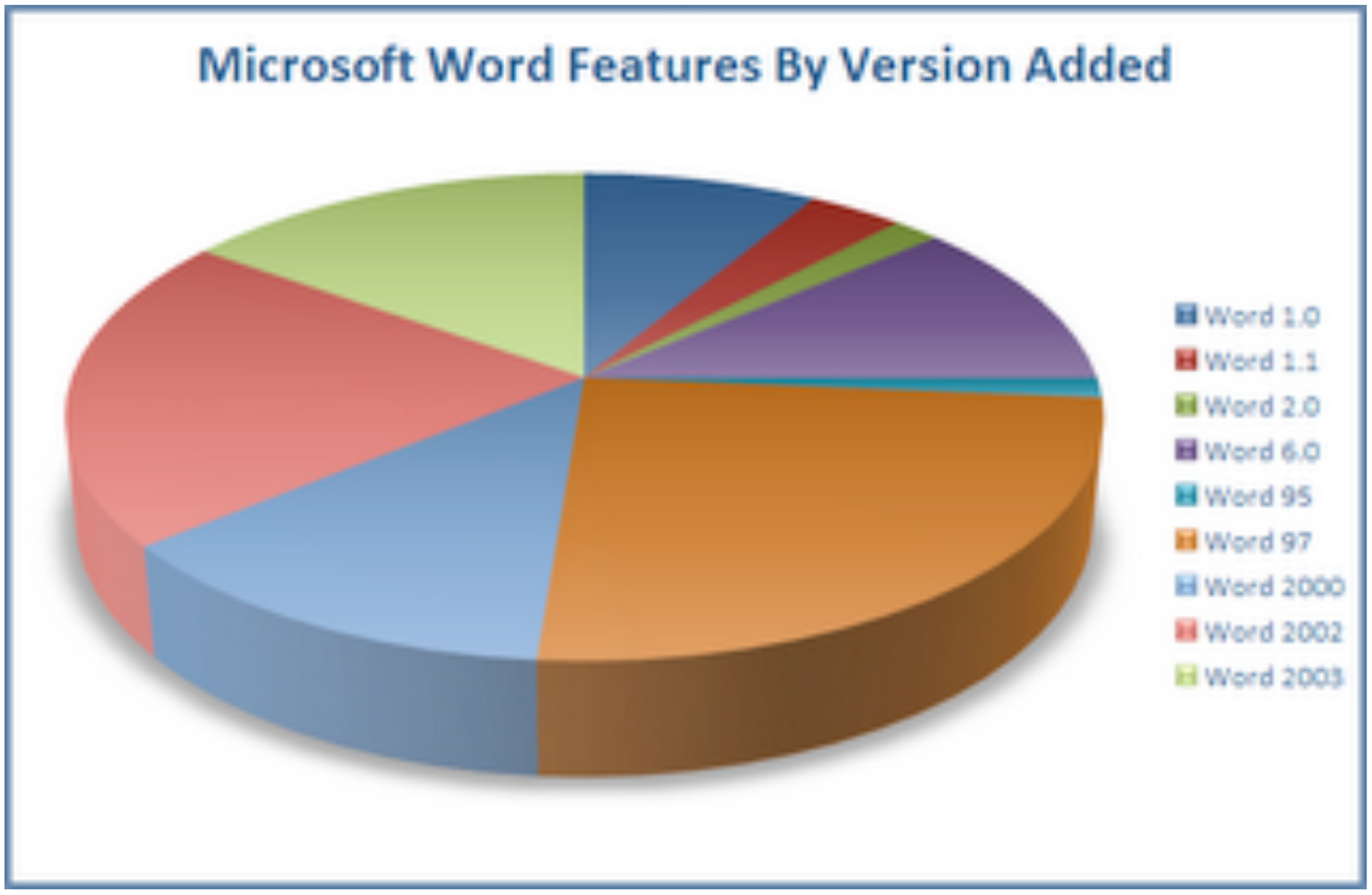
55%
GLI INTERVISTATI
PER I QUALI A DIVIDERE
IL PAESE SONO
I **PARTITI POLITICI**

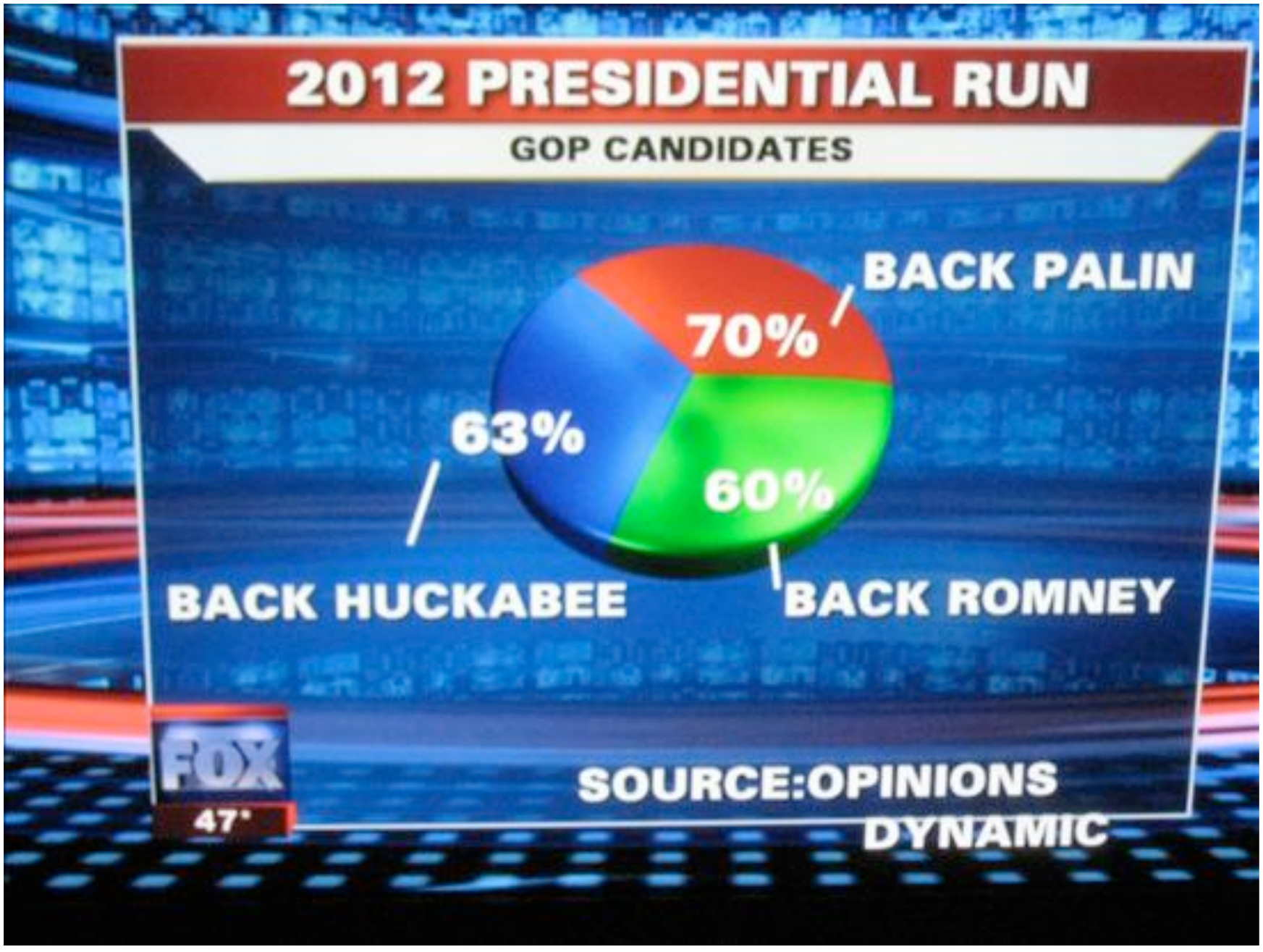
40%
GLI INTERVISTATI CHE
CONSIDERANO **LA MAFIA**
IL NOSTRO MAGGIOR
MOTIVO DI VERGOGNA

30%
GLI INTERVISTATI
SECONDO I QUALI **L'ARTE**
DI ARRANGIARSI È LA NOSTRA
MIGLIORE QUALITÀ



The Archisvist, a Microsoft Desktop app to archive tweets


Microsoft Word Features By Version Added
Word 1.0
Word 1.1
Word 2.0
Word 6.0
Word 95
Word 97
Word 2000
Word 2002
Word 2003



FOX News, November 26 2009

misurare

measuring

a **process**

assigns numbers to things

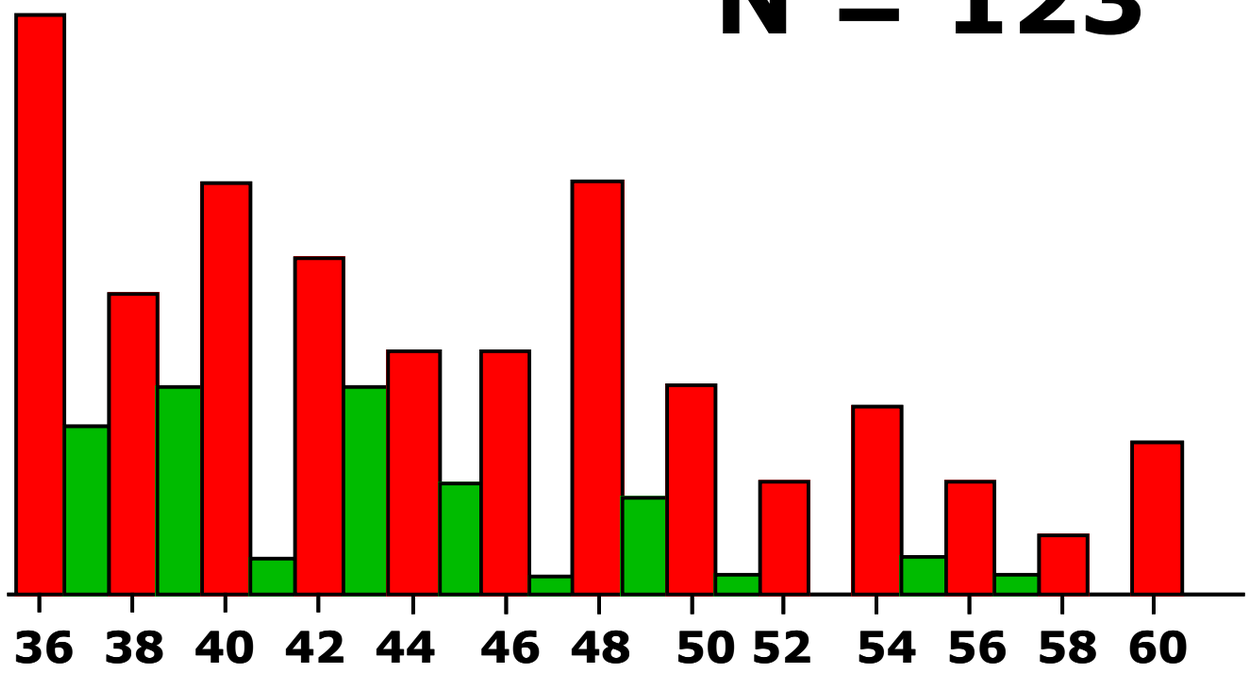
relations between things must correspond to relations between numbers

	things	numbers
A	———	10
B	———————	20

relations		
	B longer than A	20 > 10
	L(A conc. B) = L (A) & L(B)	30 = 10 + 20

to understand the measure, you must understand the process

Italian school grades
1993/94



distinzione utile

validità

cosa stiamo misurando?

il “significato”

affidabilità

stiamo misurando bene?

la “replicabilità”

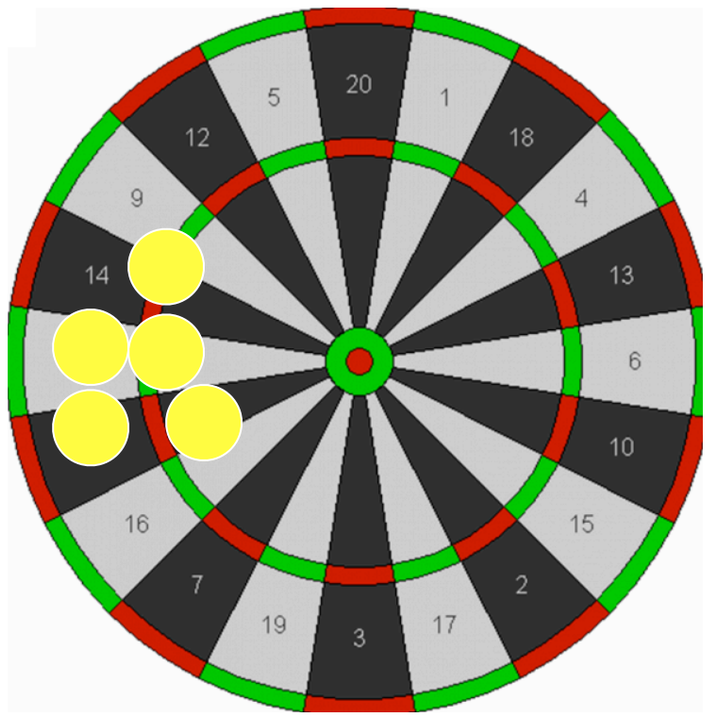
errors

variable error (random variation)

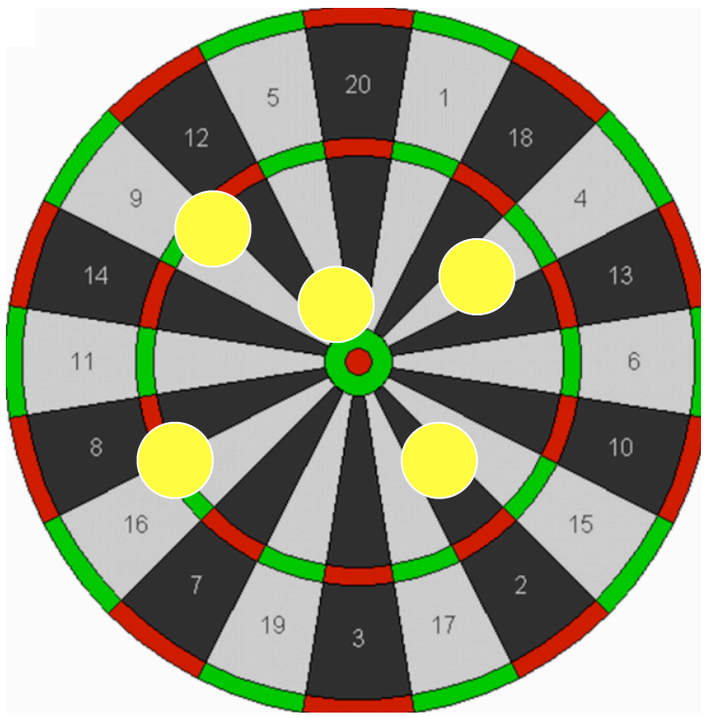
constant error (systematic bias)

precision: inverse of VE

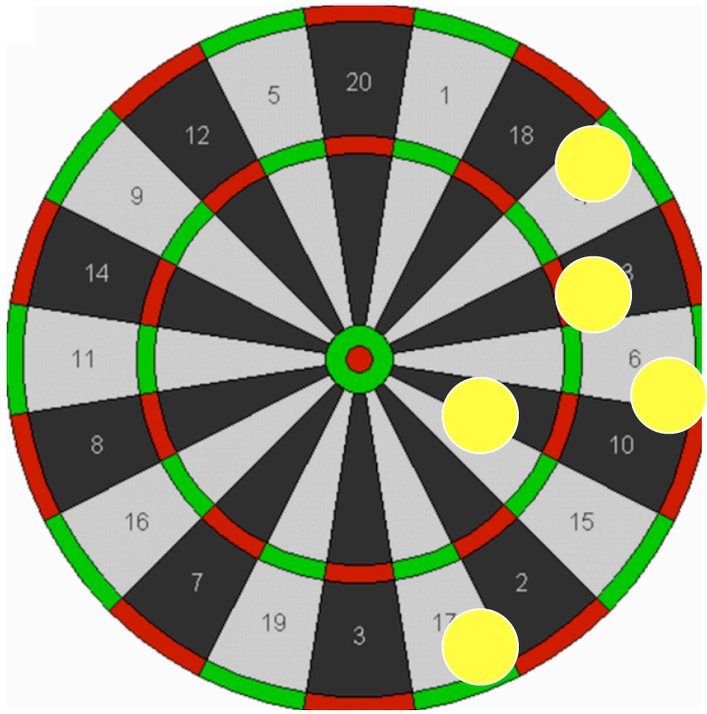
accuracy: inverse of CE



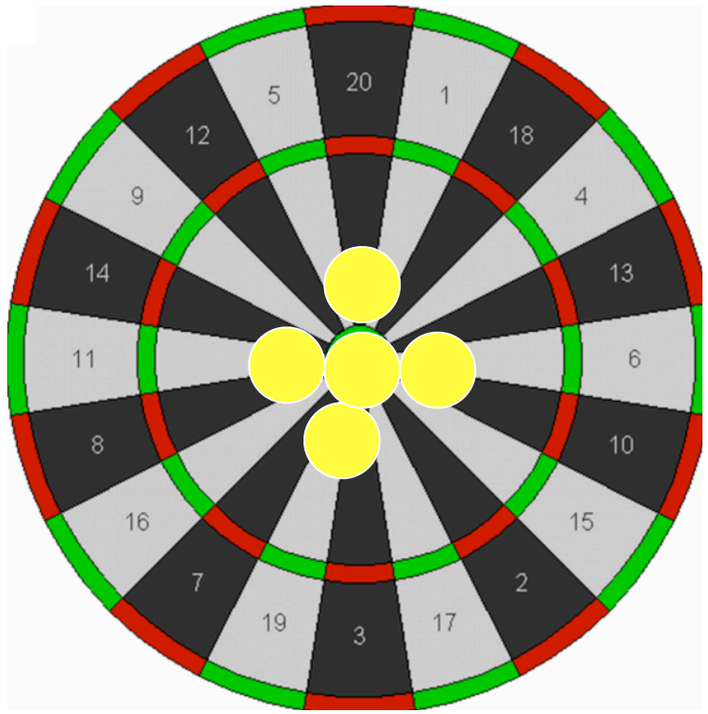
high precision
low accuracy



low precision
high accuracy



low precision
low accuracy



high precision
high accuracy

R

simulare una serie di misure
la funzione rnorm()
la funzione hist()

scale di misura e tipologie di dati

Stanley S. Stevens
(1906 - 1973)

Stevens, S. S. (1946).
On the Theory of Scales of Measurement.
Science, **103** (2684), 677–680.

Scale types with their properties according to Stanley Smith Stevens					
		Nominal scale	**Ordinal** scale	**Interval** scale	**Ratio** scale
Logical/ math operations	× ÷	✗	✗	✗	✓
	+ -	✗	✗	✓	✓
	< >	✗	✓	✓	✓
	= ≠	✓	✓	✓	✓
Examples: *Dichotomous and non-dichotomous* Variable name (data values)		*Dichotomous:* Gender (male vs. female) *Non-dichotomous:* Nationality (American/Chinese/etc)	*Dichotomous:* Health (healthy vs. sick), Truth (true vs. false), Beauty (beautiful vs.ugly) *Non-dichotomous:* Opinion ('completely agree'/ 'mostly agree'/ 'mostly disagree'/ 'completely disagree')	Date (from 1457 BC to AD 2013) Latitude (from +90° to −90°)	Age (from 0 to 99 years)

wikipedia.org/wiki/Level_of_measurement

Nominal, Ordinal, Interval, and Ratio Typologies are Misleading

Paul F. Velleman
Cornell University and Data Description, Inc.
Leland Wilkinson
SYSTAT, Inc. and Northwestern University

The American Statistician (1993), 47:1, 65-72.

It is relatively easy to construct situations in which the scale type of data depends on its interpretation or on what additional information is available.

At a reception sponsored by the ASA Section on Statistical Computing and the Section on Statistical Graphics, consecutively numbered tickets, starting with “1”, were allotted at the door as people entered so that a raffle could be held.

As a winning number, 126, was selected and announced, one participant compared it to her ticket to see if she had won, thus interpreting the “126” correctly as a nominal value.

She then immediately looked around the room and remarked that “It doesn't look like there are 126 people here”, now interpreting the same value, again correctly (but using the additional information that tickets had been allotted consecutively starting with 1), as a ratio-scale value.

One of the authors compared his ticket number (56) to the winning value and realized that he had arrived too soon to win the prize, thus interpreting the values ordinally.

If additional data about the rate and regularity of arrivals had been available, he might have tried to estimate by how much longer he should have delayed his arrival from the 70-ticket difference between his ticket and the winner, thus treating the ticket number as an interval-scaled value.

A common dataset reports facts about automobiles. One of these facts is the number of cylinders in the engine. In some analyses, this is a nominal category supporting such questions as “Are there significant differences among the gas mileages of cars with 8-cylinder, 6-cylinder, and 4-cylinder engines?” Of course, these categories are clearly ordered, so ordinal-based statistics would also be appropriate. But one might also ask about the average number of cylinders in, say, U.S. cars, and wonder whether this average had declined in recent years. This requires us to consider these data values (all of them integers) as interval-scale values — which they can certainly be. Finally, we might consider the size of each cylinder and compute the ratio of each car’s displacement to the number of its cylinders — a completely appropriate operation (for ratio- scale data).

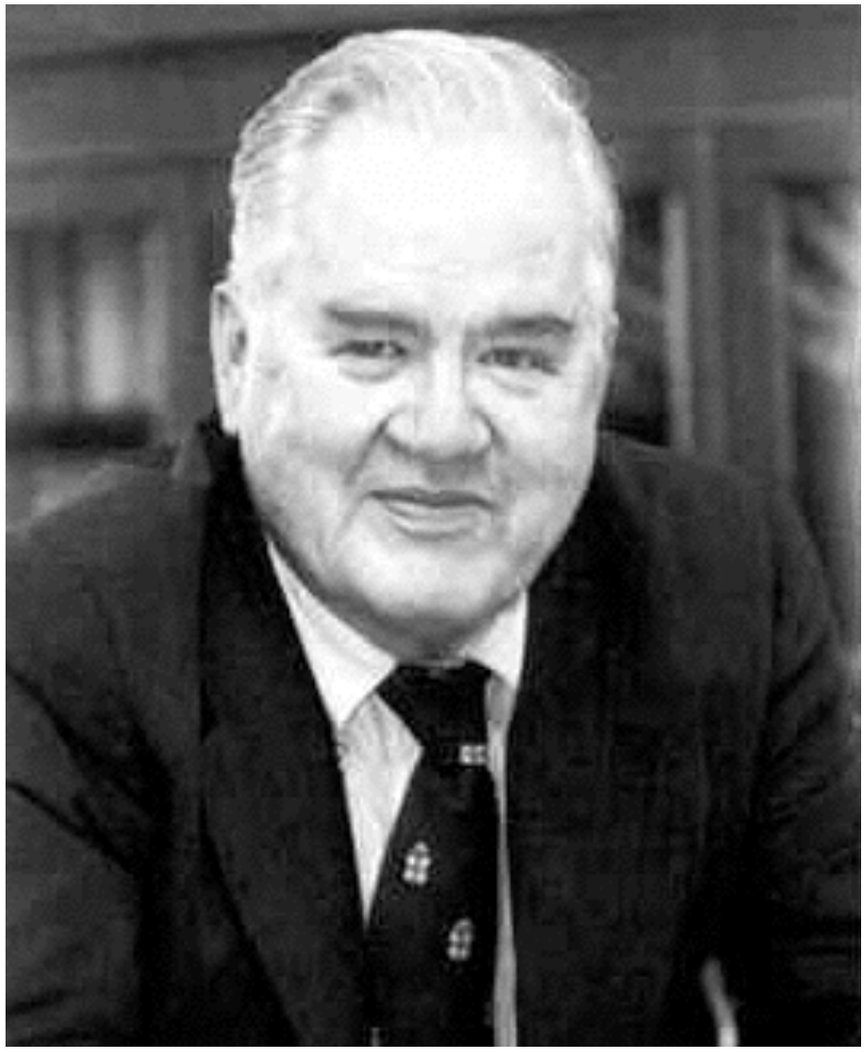
The point of these examples, of course, is that the assertion, common to many traditional statistics texts, that "data values are nominal, ordinal, interval, or ratio" simplifies the matter so far as to be false. Scale type, as defined by Stevens, is not an attribute of the data, but rather depends upon the questions we intend to ask of the data and upon any additional information we may have. It may change due to transformation of the data, it may change with the addition of new information that helps us to interpret the data differently, or it may change simply because of the questions we choose to ask.

Left								*Right*	Row Sum	
			1 1 1 1						4	*More*
		1 1 1 0				0 1 1 1			3	
	1 1 0 0			0 1 1 0			0 0 1 1		2	
1 0 0 0		0 1 0 0			0 0 1 0			0 0 0 1	1	
			0 0 0 0						0	*Less*

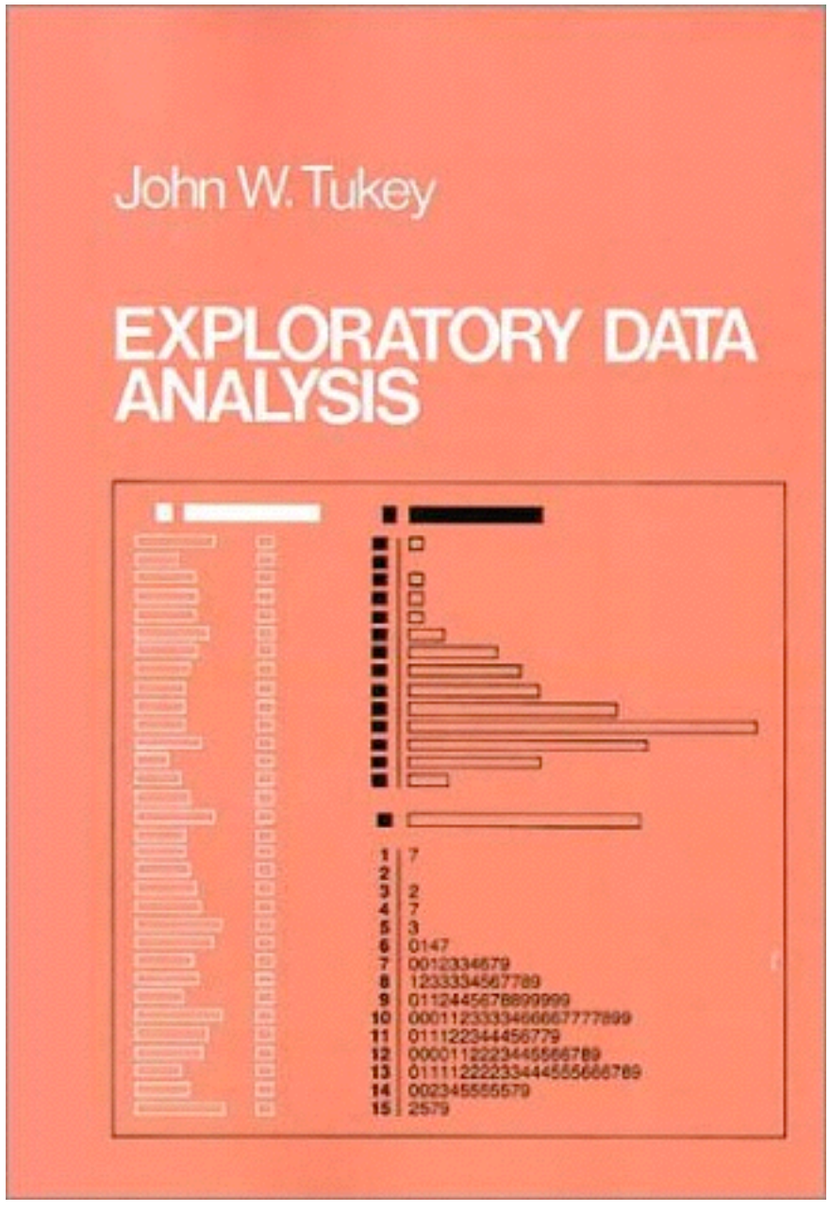
In this scale, the horizontal dimension comprises a qualitative (nominal) scale of attributes and the vertical dimension measures a quantitative (ordinal, interval, or ratio) scale. For example, each profile might be the presence or absence of each of four symptoms in a patient. In this case, the vertical scale might be related to severity of illness and the horizontal scale might be related to different syndromes.

Conclusion

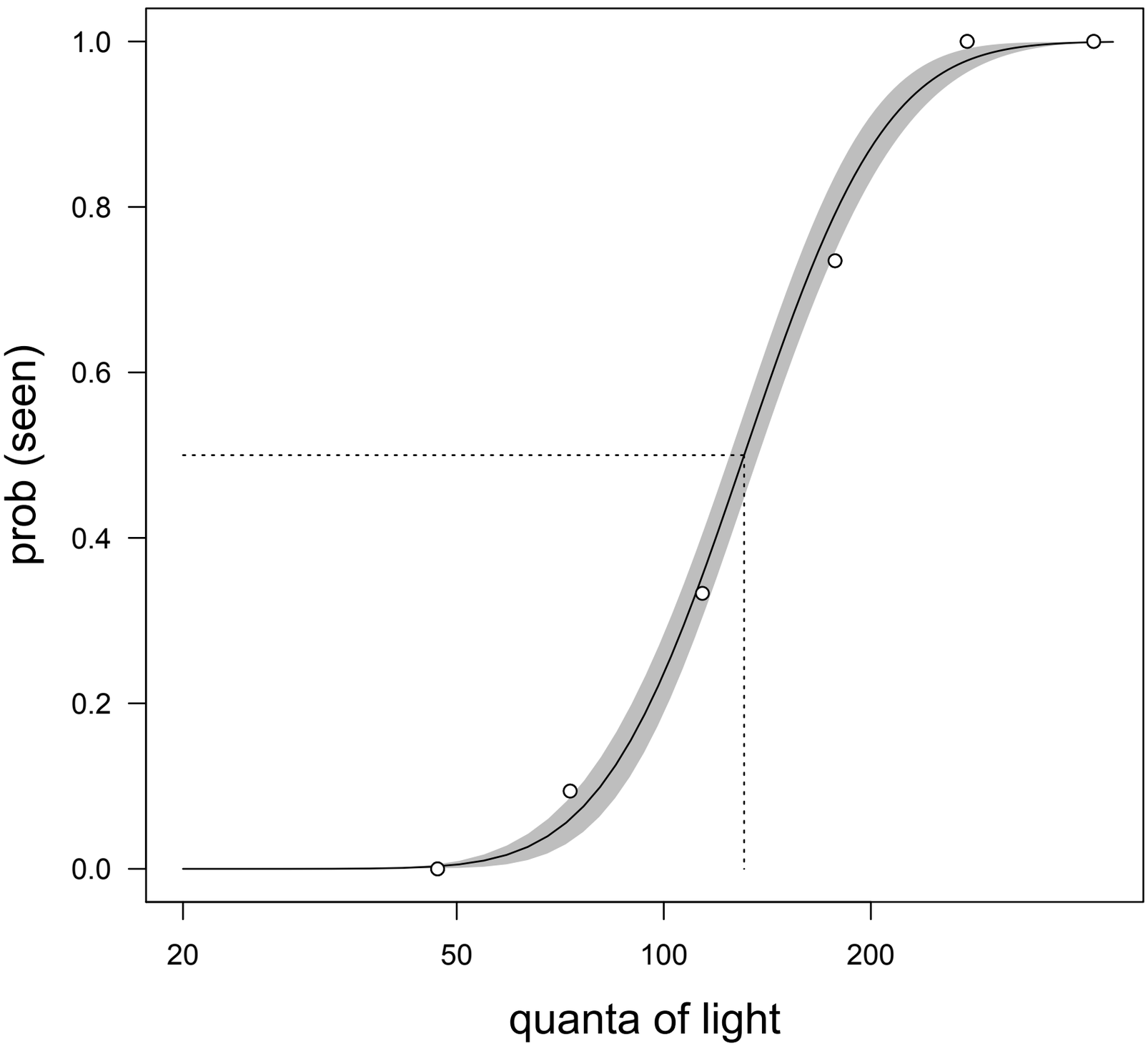
Measurement theory is important to the interpretation of statistical analyses. However, the application of Stevens's typology to statistics raises many subtle problems. Statistics programs based on Stevens's typology suggest that doing statistics is simply a matter of declaring the scale type of data and picking a model. Worse, they assert that the scale type is evident from the data *independent of the questions asked of the data*. They thus restrict the questions that may be asked of the data. Such restrictions lead to bad data analysis and bad science.



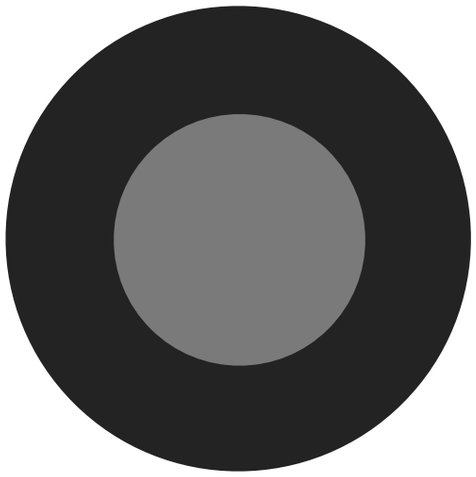
John Tukey

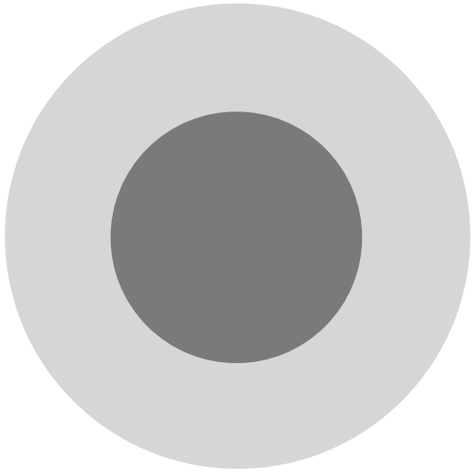

John W. Tukey
EXPLORATORY DATA ANALYSIS
1 | 7
2 |
3 | 2
4 | 7
5 | 3
6 | 0147
7 | 0012334679
8 | 1233334567789
9 | 0112445678899999
10 | 000112333346666777789
11 | 011122344456779
12 | 00001122234455566789
13 | 011112222334445555666789
14 | 002345555579
15 | 2579

tipologia	caratteristiche
nomi	etichette
gradazioni	categorie ordinate
ranghi	classifica 1 - n
frazioni *counted*	proporzioni 0 -1 o %
conteggi	interi non negativi
quantità	reali non negativi
bilanci	reali negativi e positivi

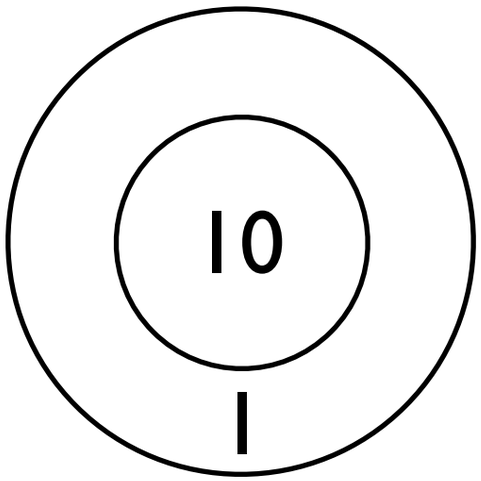

observer S.H.
1.0
0.8
0.6
0.4
0.2
0.0
prob (seen)
20
50
100
200
quanta of light

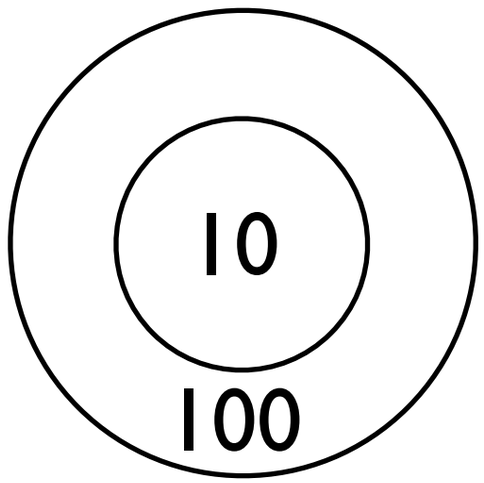
configurazioni disco-anello





luminanze





rapporti fra luminanze

10/1 = 10

10/100 = 0.1

bilancio: distanze

log10(10) = 1

log10(0.1) = -1