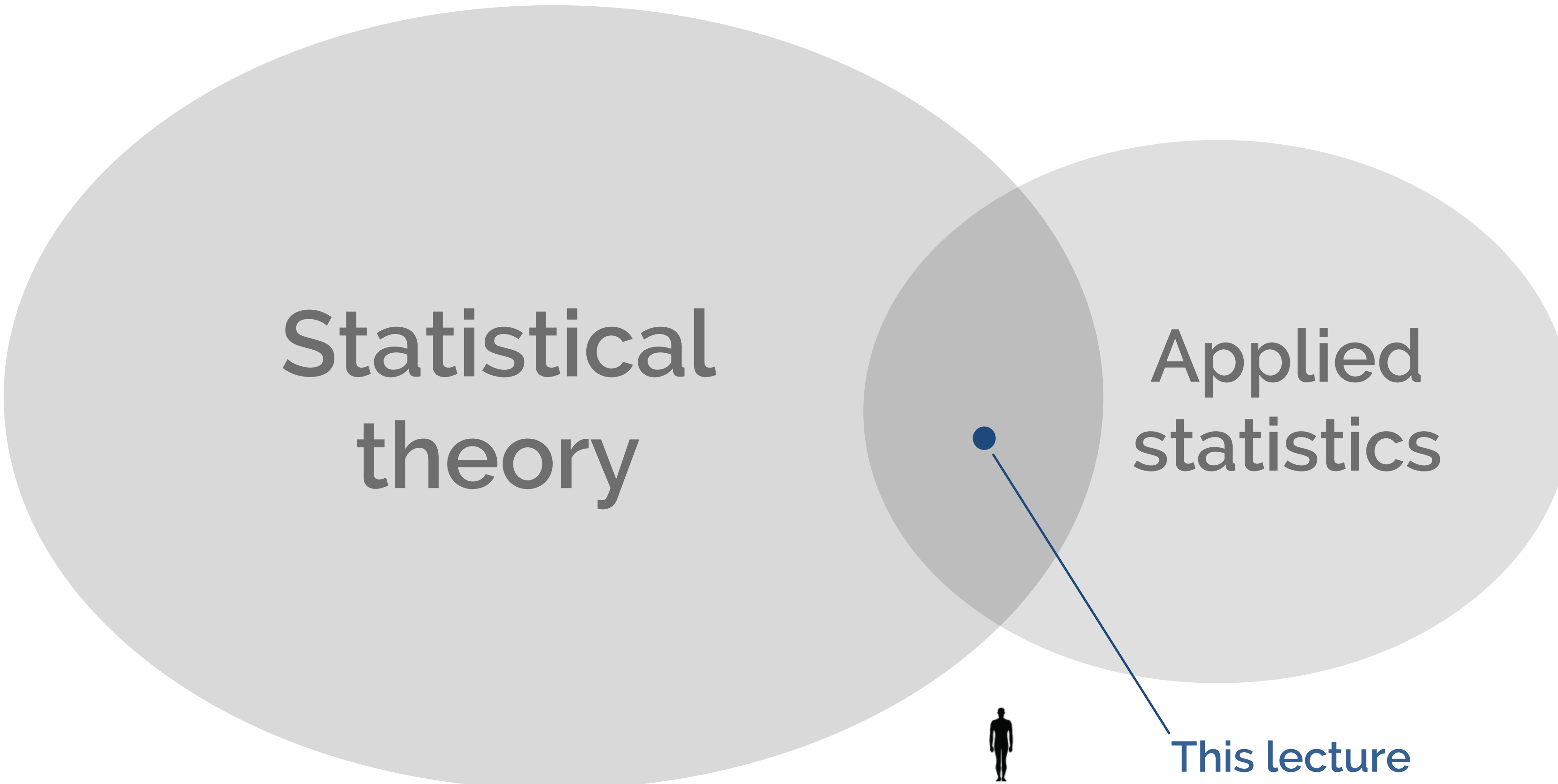
INTRODUCTION TO STATISTICS

Slides by **Pierre Dragicevic**

WHAT YOU WILL LEARN



GOALS

- Learn basic intuitions and terminology
- Perform basic statistical inference with R
- Focus on high-level aspects
- Accent on estimation rather than hypothesis testing ("the New Statistics")

ORGANIZATION

- Part I - Elementary notions
- Part II - Tutorial with R
- Part III - Assignments

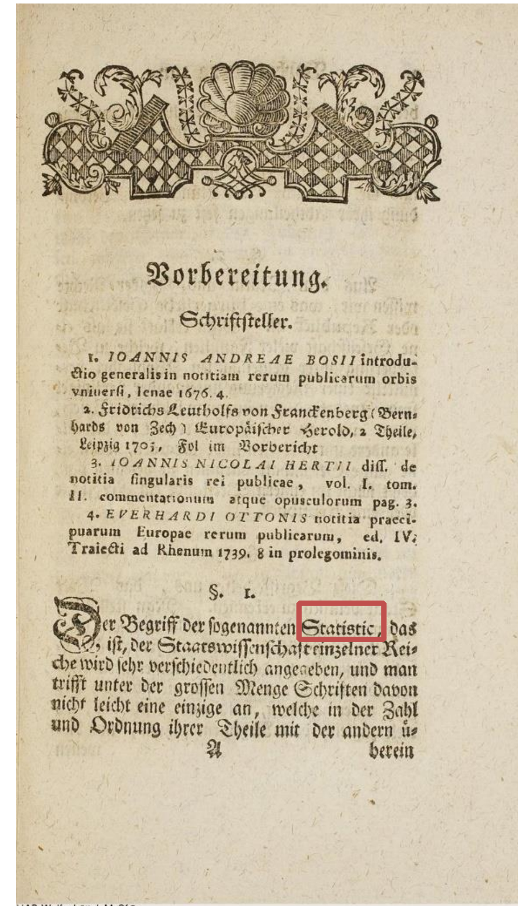
A DEFINITION

- Statistics is the study of the collection, analysis, interpretation, presentation and organization of data.

Dodge, Y. (2006) The Oxford Dictionary of Statistical Terms, OUP.

ORIGINS

- **1750s German Statistik**
  "analysis of data about the state"
- **Quickly adopted in England** (previously called "*political arithmetics*")



Vorbereitung.

Schriftsteller.

1. *IOANNIS ANDREAE BOSII* introductio generalis in notitiam rerum publicarum orbis vniuersi, Ienae 1676. 4.
2. Fridrichs Leutholfs von Franckenberg (Bernhards von Zech) Europäischer Herold, 2 Theile, Leipzig 1705, Fol im Vorbericht
3. *IOANNIS NICOLAI HERTII* diss. de notitia singularis rei publicae, vol. I. tom. II. commentationum atque opusculorum pag. 3.
4. *EVERHARDI OTTONIS* notitia praecipuarum Europae rerum publicarum, ed. IV. Traiecti ad Rhenum 1739. 8 in prolegominis.

§. 1.

Der Begriff der sogenannten Statistic, das ist, der Staatswissenschaft einzelner Reiche wird sehr verschiedentlich angegeben, und man trifft unter der grossen Menge Schriften davon nicht leicht eine einzige an, welche in der Zahl und Ordnung ihrer Theile mit der andern übereinA herein

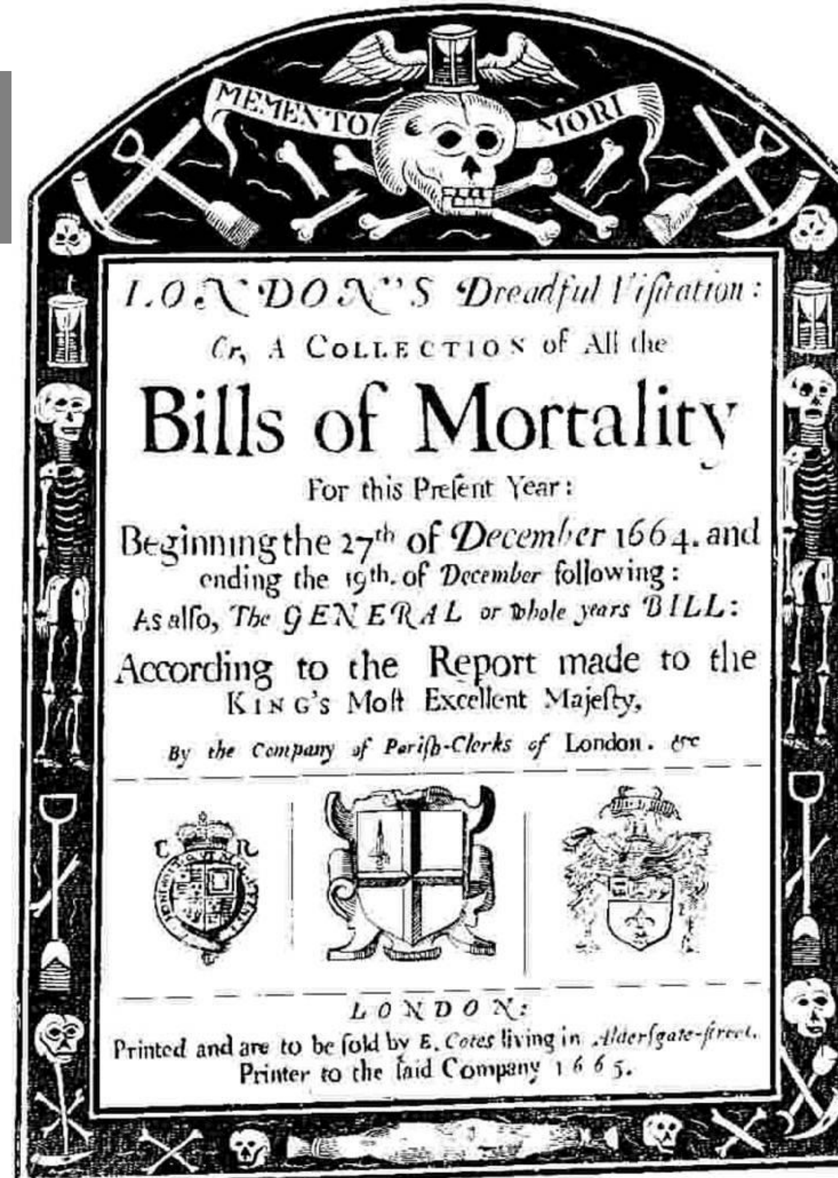
http://www.deutschestextarchiv.de/book/view/achenwall_staatswissenschaft_1749/?hl=Statistic&p=15

Achenwall, Gottfried: Abriß der neuesten Staatswissenschaft der vornehmsten Europäischen Reiche und Republicken. Göttingen, 1749

ORIGINS

- John Graunt, 1662
  Observations on the bills of mortality

CAPTAIN JOHN GRAUNT



THE TABLE OF CASUALTIES.

The Years of our Lord	1647	1648	1649	1650	1651	1652	1653	1654	1655	1656	1657	1658	1659	1660	1629	1630	1631	1632	1633	1634	1635	1636	1629 1630 1631 1632	1633 1634 1635 1636	1647 1648 1649 1650	1651 1652 1653 1654	1655 1656 1657 1658	1629 1649 1659	In 20 Years.
Abortive, and stilborn	335	329	327	351	389	381	384	433	483	419	463	467	421	544	499	439	410	445	500	475	507	523	1793	2005	1342	1587	1832	1247	8559
Aged	916	835	889	696	780	834	864	974	743	892	869	1176	909	1095	579	712	661	671	704	623	794	714	2475	2814	3336	3452	3680	2377	15757
Ague, and Fever	1260	884	751	970	1038	1212	1282	1371	689	875	999	1800	2303	2148	956	1091	1115	1108	951	1279	1622	2360	4418	6235	3865	4903	4363	4010	23784
Apoplex, and sodainly	68	74	64	74	106	111	118	86	92	102	113	138	91	67	22	36		17	24	35	26		75	85	280	421	445	177	1306
Bleach			1	3	7	2				1															4	9	1	1	15
Blasted	4	1			6	6			4		5	5	3	8	13	8	10	13	6	4		4	54	14	5	12	14	16	99
Bleeding	3	2	5	1	3	4	3	2	7	3	5	4	7	2	5	2	5	4	4	3			16	7	11	12	19	17	65
Bloudy Flux, Scouring, and Flux	155	176	802	289	833	762	200	386	168	368	362	233	346	251	449	438	352	348	278	512	346	330	1587	1466	1422	2181	1161	1597	7818
Burnt, and Scalded	3	6	10	5	11	8	5	7	10	5	7	4	6	6	3	10	7	5	1	3	12	3	25	19	24	31	26	19	125
Calenture	1			1		2	1	1			3										1	3		4	2	4	3		13
Cancer, Gangrene, and Fistula	26	29	31	19	31	53	36	37	73	31	24	35	63	52	20	14	23	28	27	30	24	30	85	112	105	157	150	114	609
Wolf				8																				8					8
Canker, Sore-mouth, and Thrush	66	28	54	42	68	51	53	72	44	81	19	27	73	68	6	4	4	1			5	74	15	79	190	244	161	133	689
Childbed	161	106	114	117	206	213	158	192	177	201	236	225	226	194	150	157	112	171	132	143	163	230	590	608	498	769	839	490	3364
Chrisomes, and Infants	1369	1254	1065	990	1237	1280	1050	1343	1089	1393	1162	1144	858	1123	2596	2378	2035	2268	2130	2315	2113	1895	9277	8453	4678	4910	4788	4519	32106
Colick, and Wind	103	71	85	82	76	102	[illegible]	101	85	120	113	179	116	167	48	57					37	50	105	87	341	359	497	147	1389
Cold, and Cough							41	36	21	58	30	31	33	24	10	58	51	55	45	54	50	57	174	207	00	77	140	43	598
Consumption, and Cough	2423	2200	2388	1988	2350	2410	2286	2868	2606	3184	2757	3610	2982	3414	1827	1910	1713	1797	1754	1955	2080	2477	5157	8266	8999	9914	12157	7197	44487
Convulsion	684	491	530	493	569	653	606	828	702	1027	807	841	742	1031	52	87	18	241	221	386	418	709	498	1734	2198	2656	3377	1324	9073
Cramp			1														1	0	0	0	0	0	01	00	01	0	0	1	2
Cut of the Stone		2	1	3		1	1	2	4	1	3	5	46	48				5	1	5	2	2	5	10	6	4	13	47	38
Dropsy, and Tympany	185	434	421	508	444	556	617	704	660	706	631	931	646	872	235	252	279	280	266	250	329	389	1048	1734	1538	2321	2982	1302	9623
Drowned	47	40	30	27	49	50	3	30	43	4	63	60	57	48	43	33	29	14	37	32	32	45	139	147	144	182	215	130	827
Excessive drinking			2																						2			2	2
Executed	8	17	29	43	24	12	19	21	19	22	20	18	7	18	19	13	12	18	13	13	13	13	62	52	97	76	79	55	384
Fainted in a Bath					1																					1			1
Falling-Sickness	3	2	2	3		3	4	1	4	3	1		4	5	3	10	7	7	2	5	6	8	27	21	10	8	8	9	74
Flox, and small Pox	139	400	1190	184	525	1279	139	812	1294	823	835	409	1523	354	72	40	58	531	72	1354	293	127	701	1840	1913	2755	3361	2785	10576
Found dead in the Streets	6	6	9	8	7	9	14	4	3	4	9	11	2	6	18	33	20	6	13	8	24	24	83	69	[illegible]	34	27	29	243
French-Pox	18	29	15	18	21	20	10	20	29	23	25	53	51	31	17	12	12	12	7	17	12	22	53	48	80	81	130	83	392
Frighted	4	4	1		3		2		1	1				9	1			1				3	2	3	9	5	2	2	21
Gout	9	5	12	9	7	7	5	6	8	7	8	13	14	2	2	5	3	4	4	5	7	8	14	24	35	25	36	28	134
Grief	12	13	16	7	17	14	11	17	10	13	10	12	13	4	18	20	22	11	14	17	5	20	71	50	48	59	45	47	279
Hanged, and made-away themselves	11	10	13	14	9	14	13	9	14	16	24	18	11	36	8	8	6	15		3	8	7	37	18	48	47	72	32	222
[illegible]		1	1	2		2	0	6	5	3	4	5	35	26							4	2	0	6	14	14	17	46	[illegible]
Jaundice	57	35	39	49	41	43	57	71	61	41	46	77	102	76	47	59	35	43	35	45	54	63	184	197	180	212	225	188	998
Jaw-faln	1	1			3			2	2		3	1			10	16	13	8	10	10	4	11	47	35	02	5	6	10	95
Impostume	75	61	65	59	80	105	79	90	92	122	80	134	105	96	58	76	73	74	50	62	73	130	282	315	260	354	428	228	1639
Itch		1																	10				00	10	01				11
Killed by several Accidents	27	57	39	94	47	45	57	58	52	43	52	47	55	47	54	55	47	46	49	41	51	60	202	201	217	207	194	148	1021
King's Evil	27	26	22	19	22	20	26	26	27	24	23	28	28	54	16	25	18	38	35	20	26	69	97	150	94	94	102	66	537
Lethargy	3	4	2	4	4	4	3	10	9	4	6	2	6	4	1		2	2	3		2	2	5	7	13	21	21	9	67
Leprosy			1									1		2	2						2		2	2	1		1	3	[illegible]
Livergrown, Spleen, and Rickets	53	46	56	59	65	72	67	65	52	50	38	51	8	15	94	112	99	87	82	77	98	99	392	356	213	269	191	158	1421
Lunatique	12	18	6	11	7	11	9	12	6	7	13	5	14	14	6	11	6	5	4	2	2	5	28	13	47	39	31	26	[illegible]

Cancer, Gangrene, and Fistula	26	29	31	19	31	53	36	37	73	31
Wolf				8						
Canker, Sore-mouth, and Thrush	66	28	54	42	68	51	53	72	44	81
Childbed	161	106	114	117	206	213	158	192	177	201
Chrisomes, and Infants	1369	1254	1065	990	1237	1280	1050	1343	1089	1393
Colick, and Wind	103	71	85	82	76	102	[illegible]	101	85	120
Cold, and Cough							41	36	21	58
Consumption, and Cough	2423	2200	2388	1988	2350	2410	2286	2868	2606	3184
Convulsion	684	491	530	493	569	653	606	828	702	1027
Cramp			1							
Cut of the Stone		2	1	3		1	1	2	4	1
Dropsy, and Tympany	185	434	421	508	444	556	617	704	660	706
Drowned	47	40	30	27	49	50	3	30	43	4
Excessive drinking			2							
Executed	8	17	29	43	24	12	19	21	19	22
Fainted in a Bath					1					
Falling-Sickness	3	2	2	3		3	4	1	4	3
Flox, and small Pox	139	400	1190	184	525	1279	139	812	1294	823
Found dead in the Streets	6	6	9	8	7	9	14	4	3	4
French-Pox	18	29	15	18	21	20	20	20	29	23
Frighted	4	4	1		3		2		1	1
Gout	9	5	12	9	7	7	5	6	8	7
Grief	12	13	16	7	17	14	11	17	10	13

ORIGINS

John Graunt, 1662

Observations on the bills of mortality

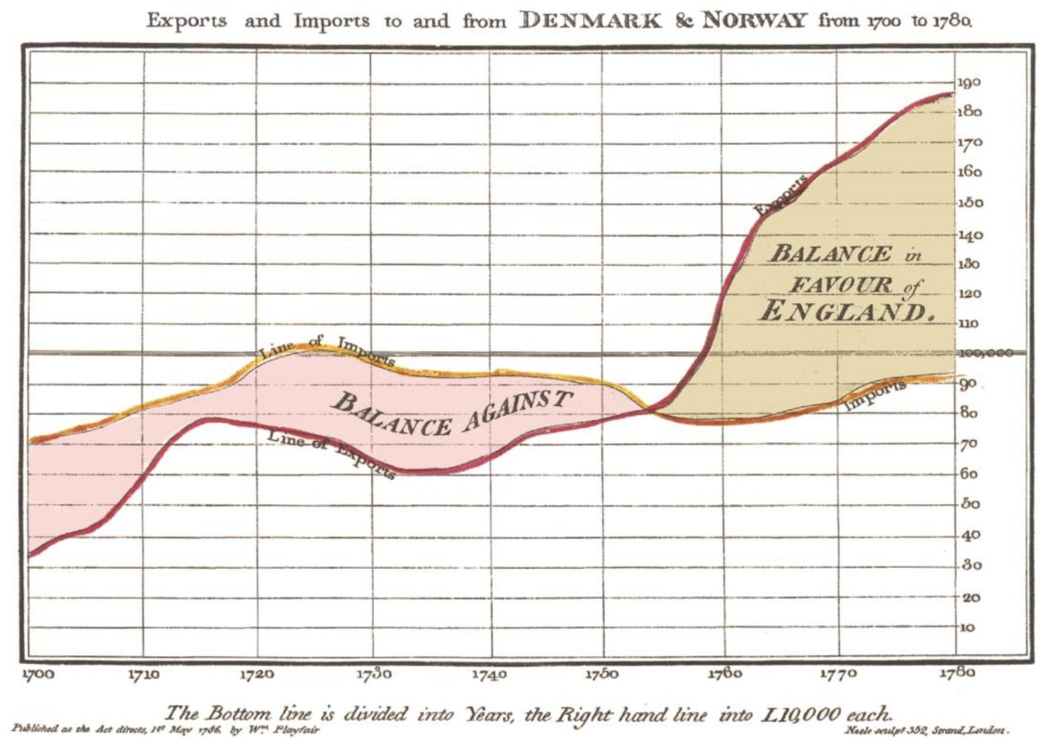
- – First “life tables”
- – Dispelled several myths about the plague
- – First analysis of sex ratio
- – First realistic estimate of the population in London

ORIGINS

- Prompted collection of more data
- Parallel developments in probability theory
- Statistics then developed into a more rigorous discipline and was applied to:
  – Business & industry
  – Medicine
  – Science
  – …

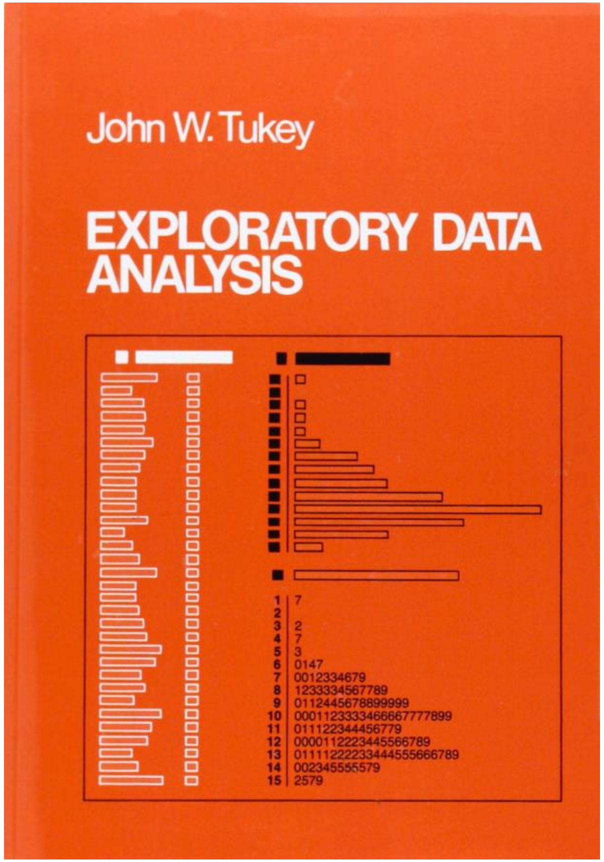
STATS & VISUALIZATION

- Statistical Charts
  - William Playfair 1759 – 1823

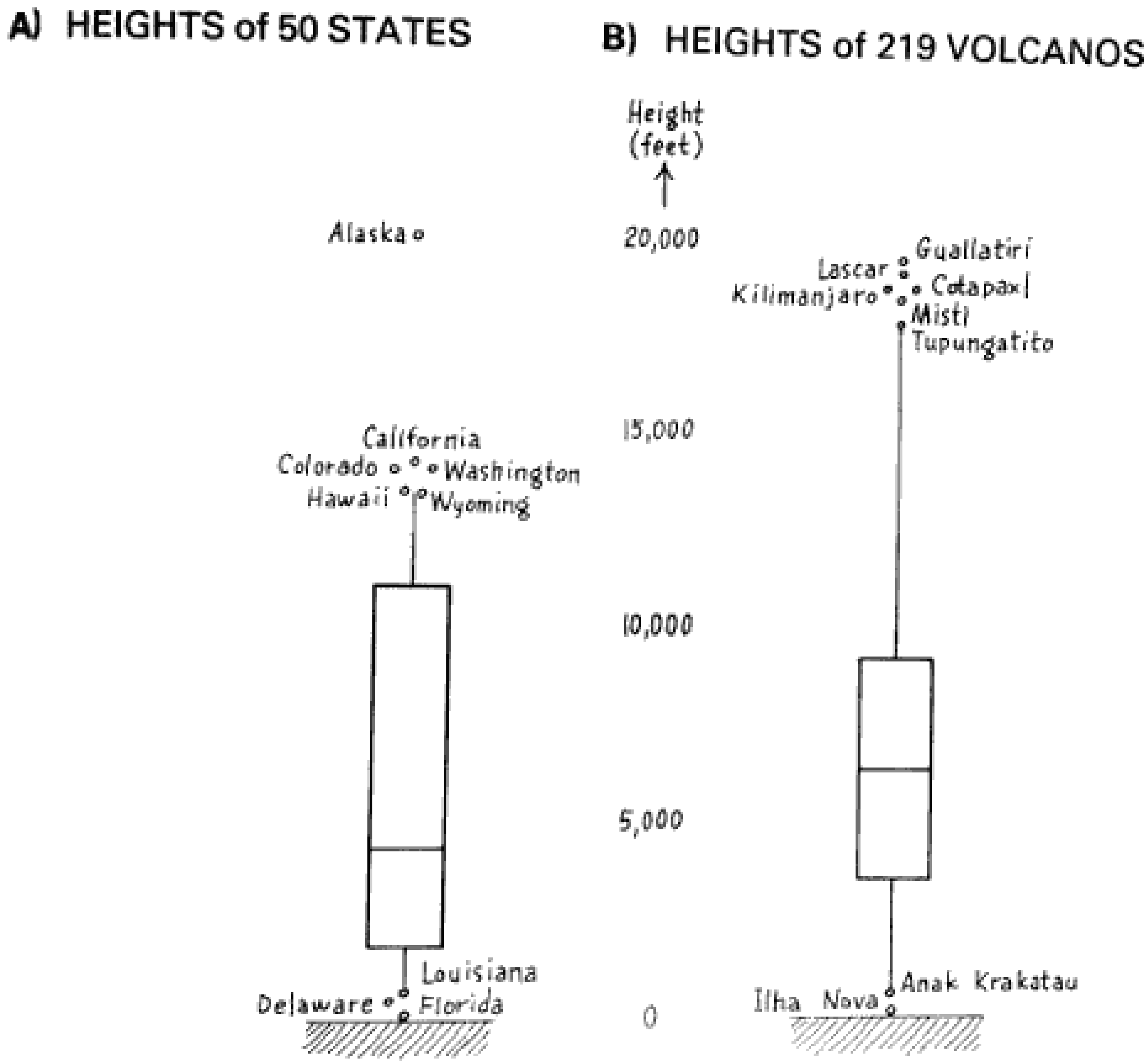


STATS & VISUALIZATION

- Exploratory Data Analysis
  - Tukey, 1977



Box-and-whisker plots with end values identified



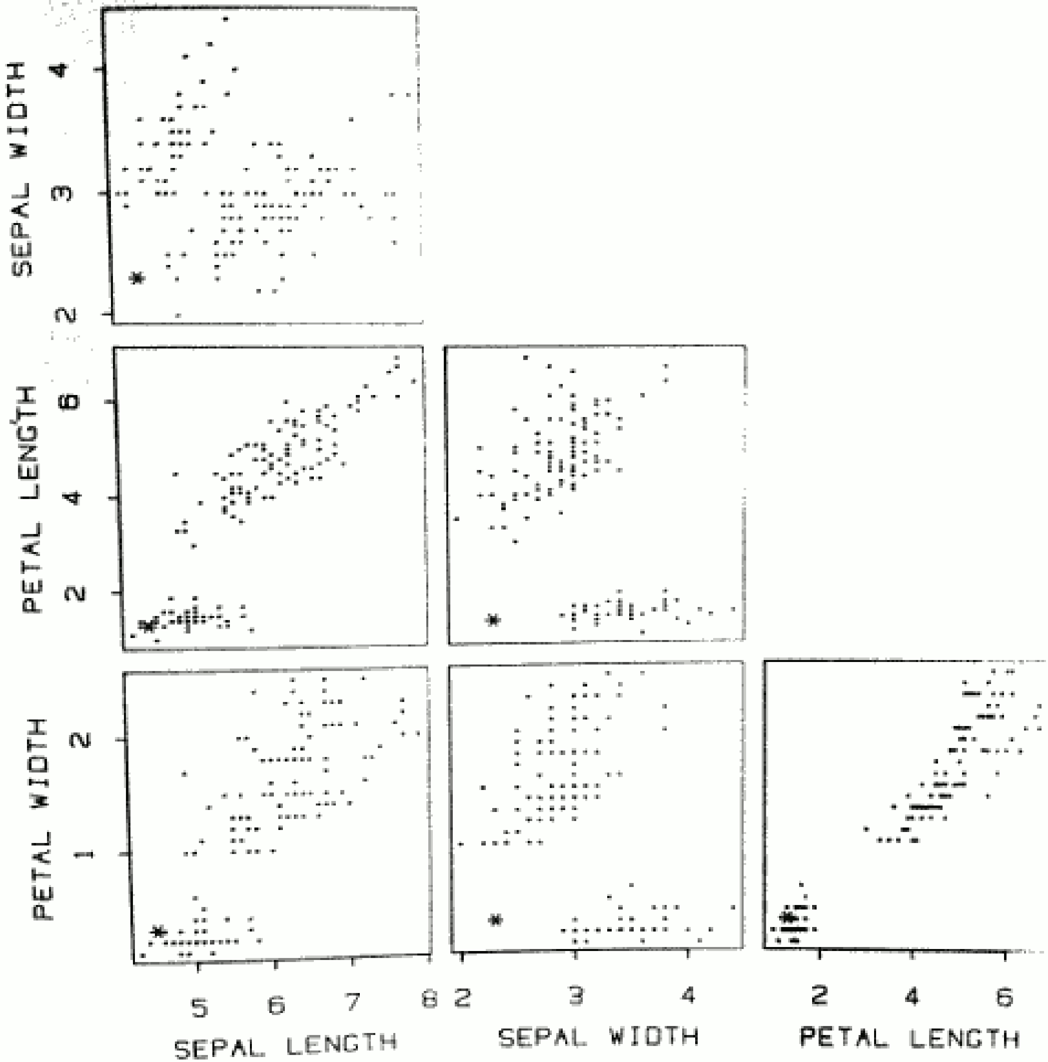
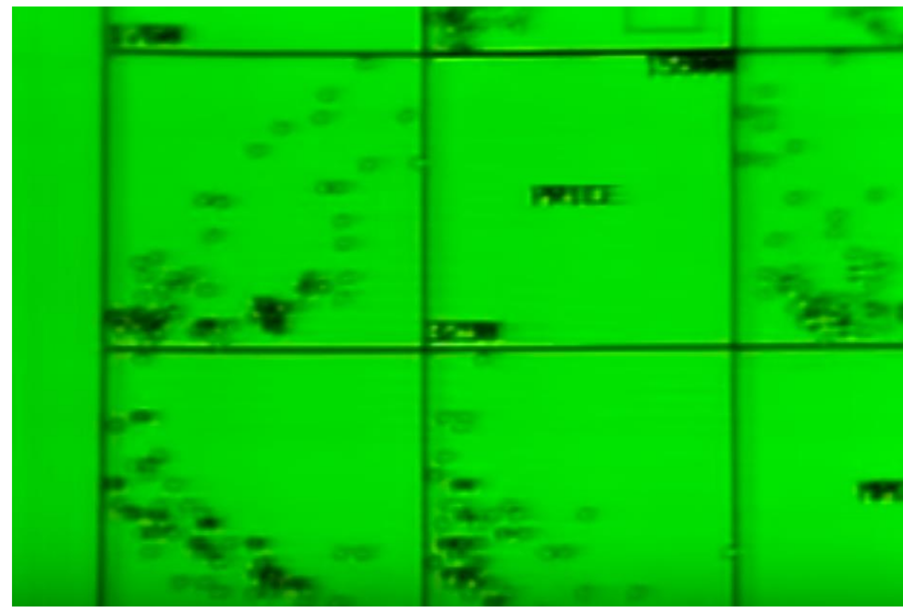
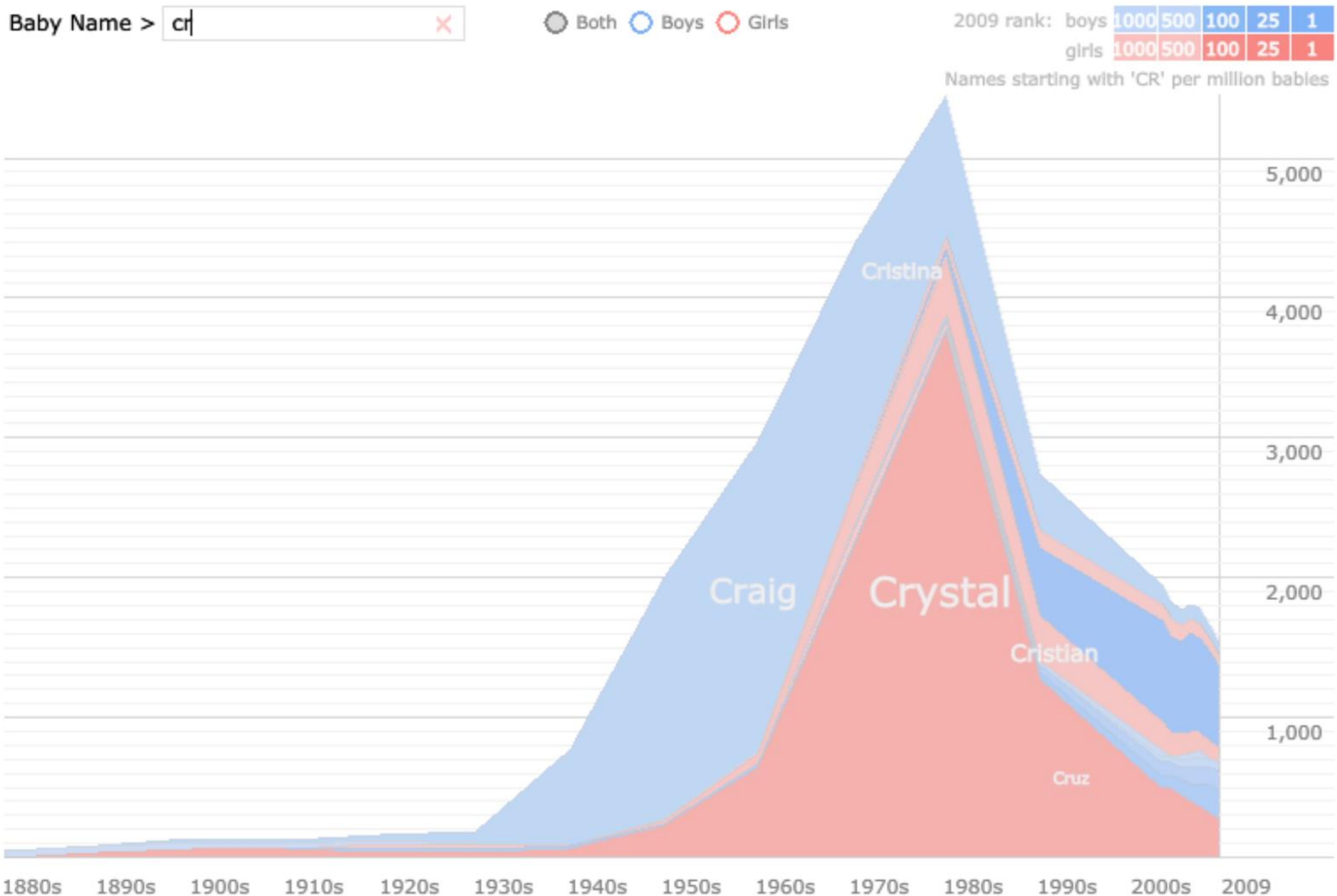


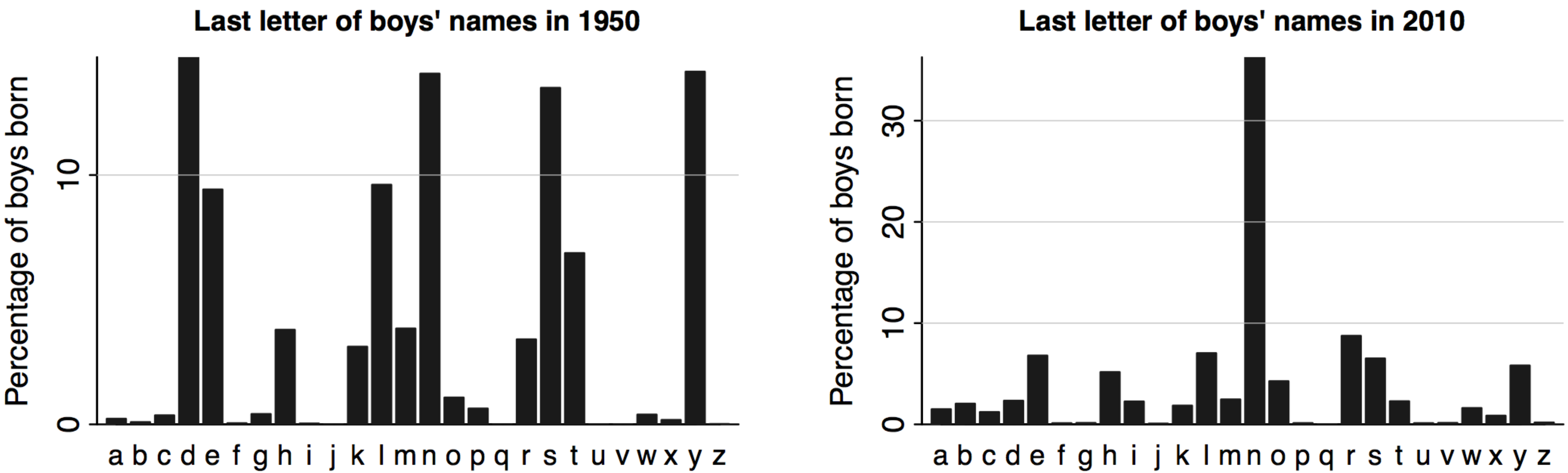
Figure 5.14 Generalized draftsman's display of the four-dimensional iris data (like Figure 5.11), with one flower plotted as an asterisk.

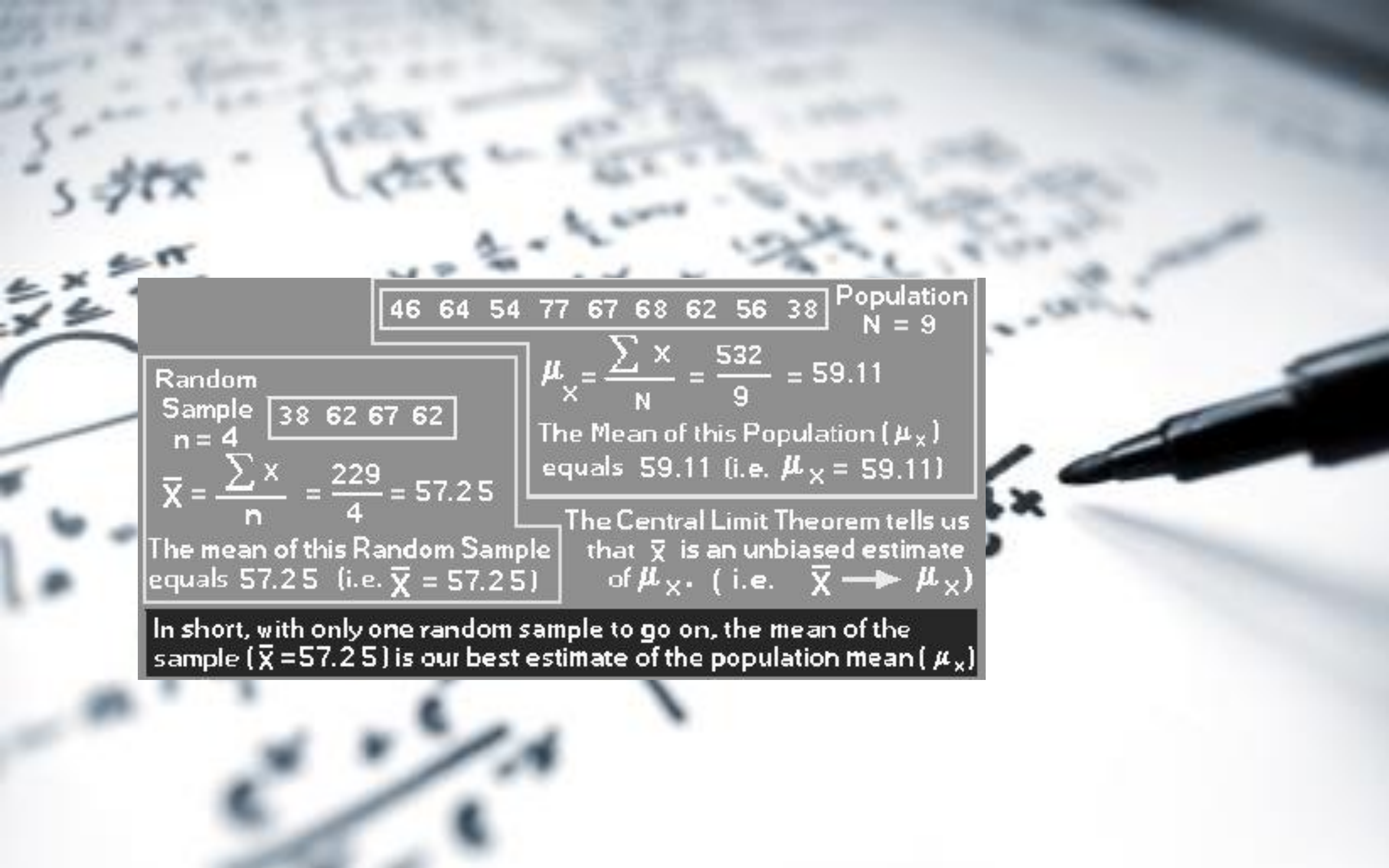
- Statistical Graphics
  - AT&T Bell Labs Video, 1985

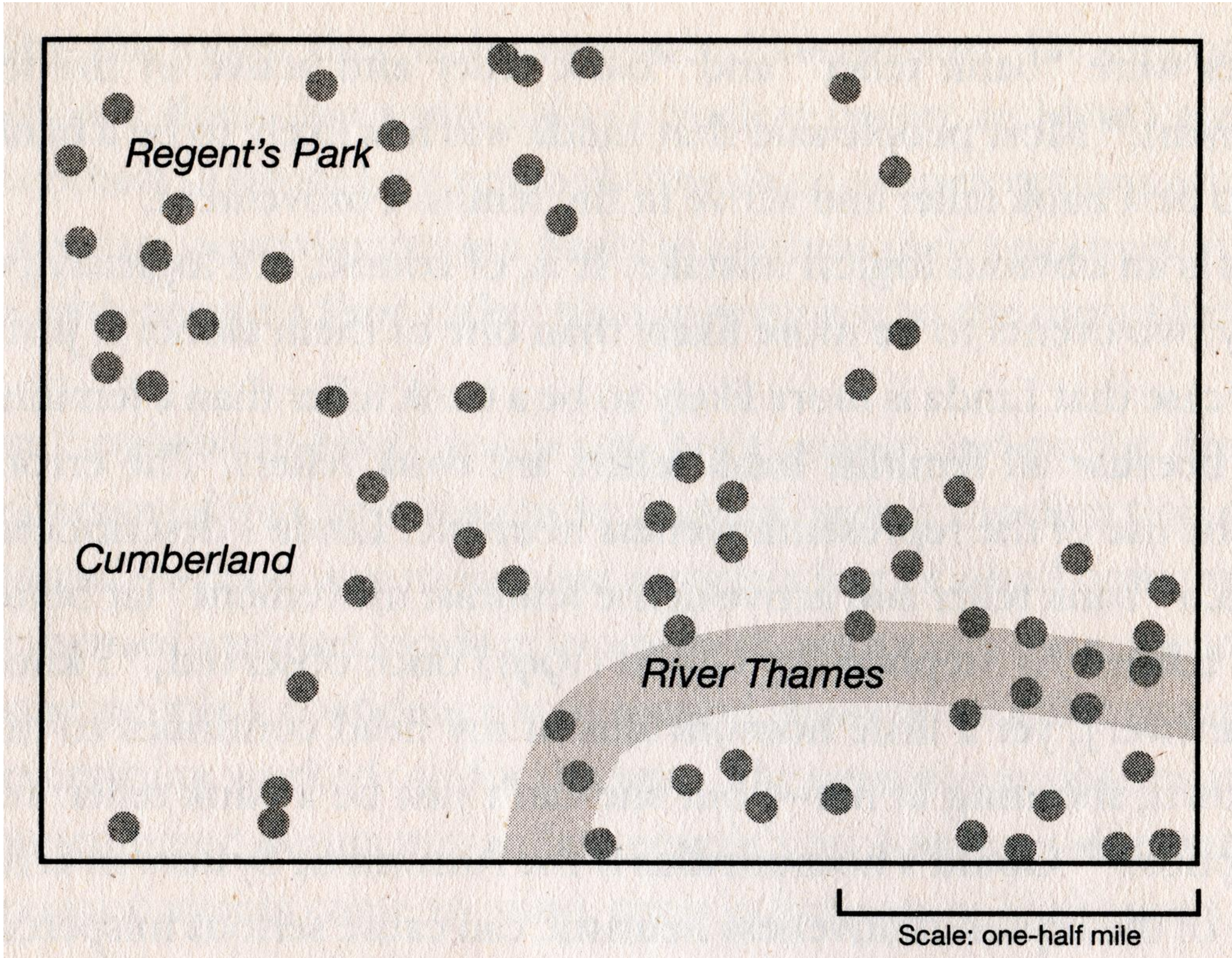




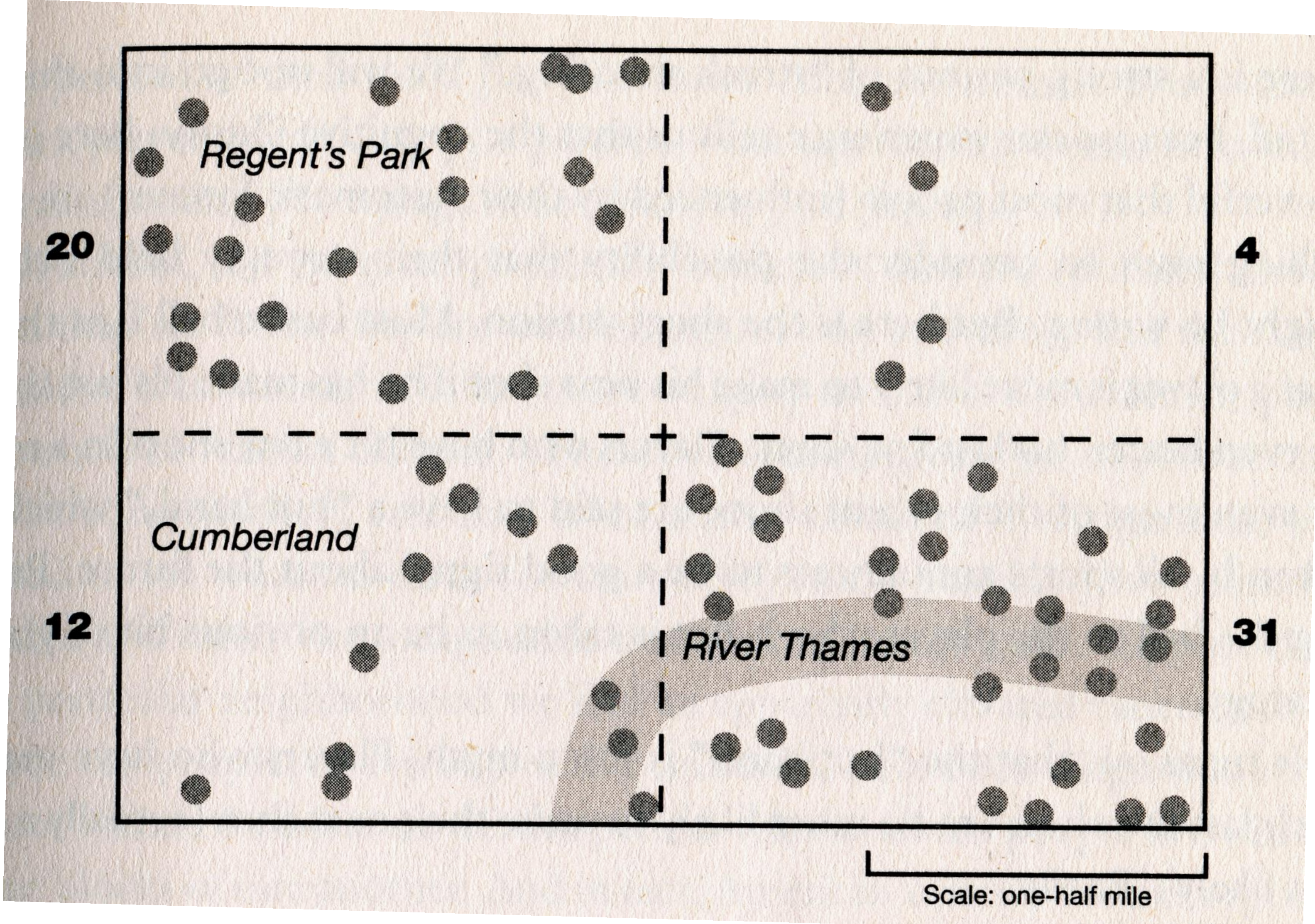
Baby Name > cr
Both
Boys
Girls
2009 rank: boys 1000 500 100 25 1
girls 1000 500 100 25 1
Names starting with 'CR' per million babies
5,000
4,000
3,000
2,000
1,000
Cristina
Craig
Crystal
Cristian
Cruz
1880s
1890s
1900s
1910s
1920s
1930s
1940s
1950s
1960s
1970s
1980s
1990s
2000s
2009


Last letter of boys' names in 1950
Percentage of boys born
0
10
a b c d e f g h i j k l m n o p q r s t u v w x y z
Last letter of boys' names in 2010
Percentage of boys born
0
10
20
30
a b c d e f g h i j k l m n o p q r s t u v w x y z

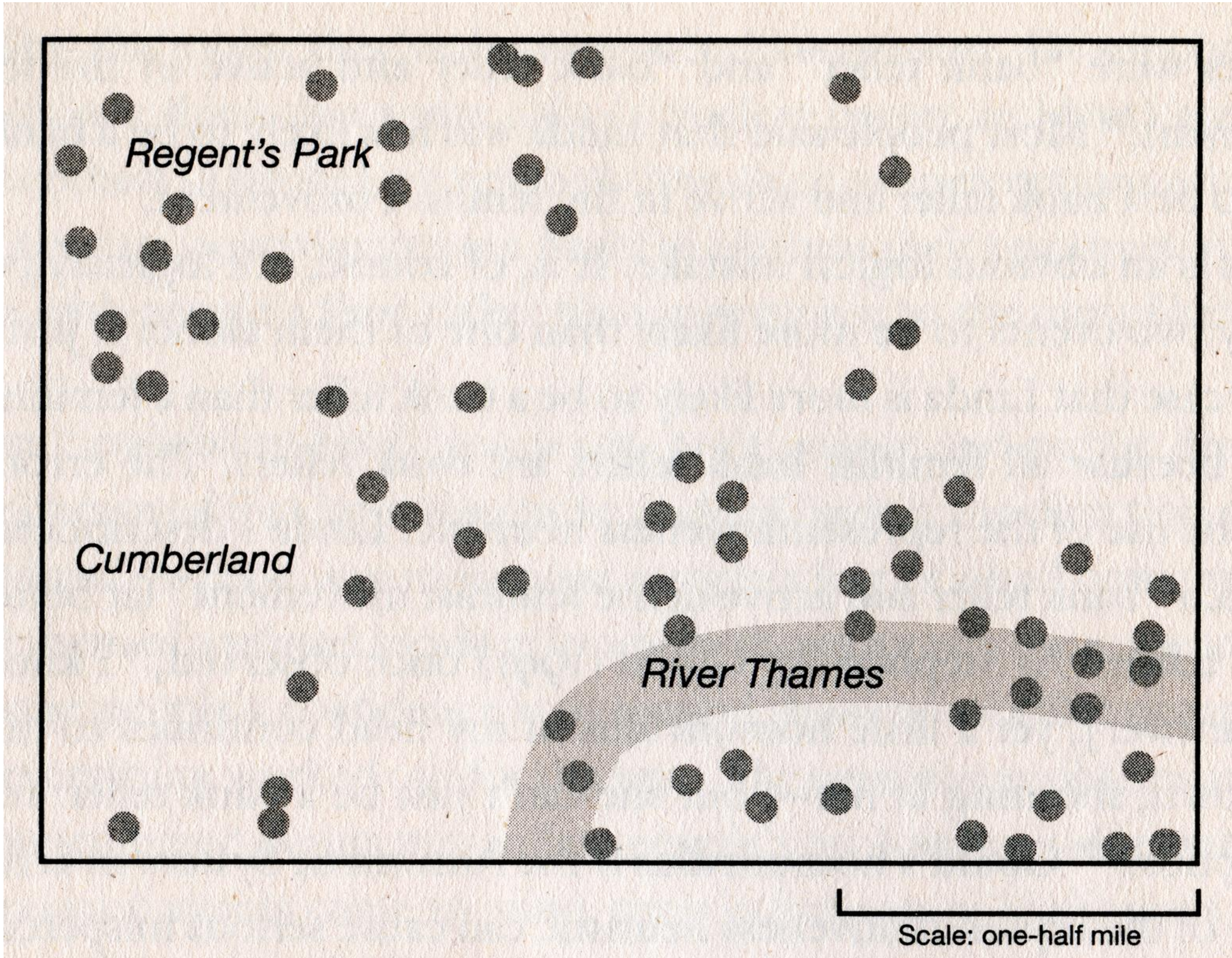

46 64 54 77 67 68 62 56 38
Population
N = 9
$\mu_X = \frac{\sum X}{N} = \frac{532}{9} = 59.11$
The Mean of this Population (μ_X) equals 59.11 (i.e. μ_X = 59.11)
Random Sample n = 4
38 62 67 62
$\bar{X} = \frac{\sum X}{n} = \frac{229}{4} = 57.25$
The mean of this Random Sample equals 57.25 (i.e. $\bar{X}$ = 57.25)
The Central Limit Theorem tells us that $\bar{X}$ is an unbiased estimate of μ_X. (i.e. $\bar{X} \rightarrow \mu_X$)
In short, with only one random sample to go on, the mean of the sample ($\bar{X}$ =57.25) is our best estimate of the population mean (μ_X)



German bombings in London during WWII



German bombings in London during WWII

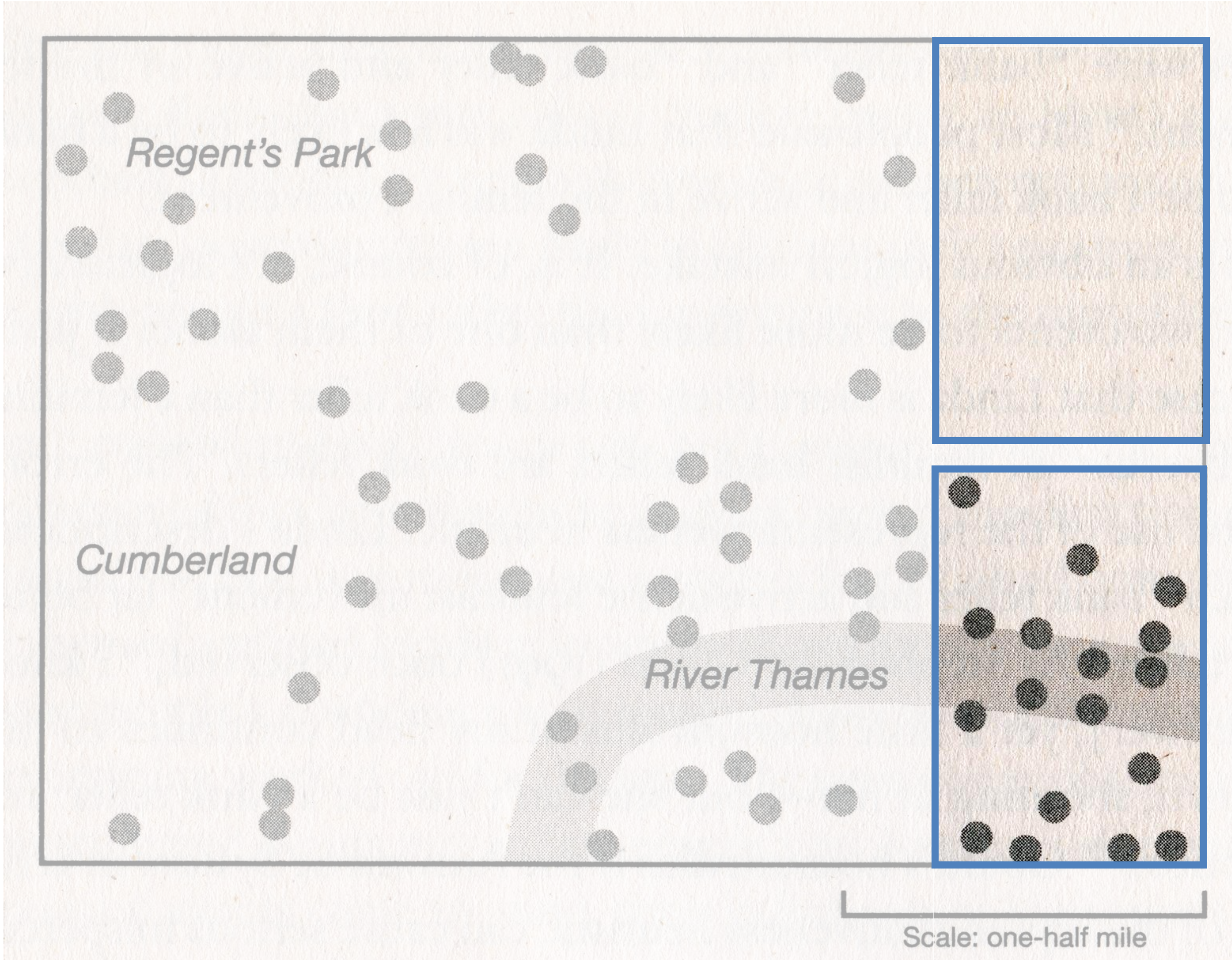


German bombings in London during WWII

STATS & VISUALIZATION

Confirmatory data analysis

- For answering questions rigorously
- Example: is this new drug effective?
- Strong focus on automatic procedures, computation and objectivity
- Looking at data can impair objectivity:
  - Cherry picking, snooping, fishing, data mining



German bombings in London during WWII

STATS & VISUALIZATION

Exploratory data analysis is sometimes compared to detective work: it is the process of gathering evidence.

Confirmatory data analysis is comparable to a court trial: it is the process of evaluating evidence.

Exploratory analysis and confirmatory analysis *"can—and should—proceed side by side"* (Tukey; 1977).

Quoted from the SAS Institute

WHAT ARE STATS?

- A set of tools and methods
- With an old tradition:
  - Origins in demographics
  - Anchored in mathematics & probability theory
  - Visual representations play a role
  - A generally strong focus on (computationally cheap) numerical calculations

WHAT ARE STATS?

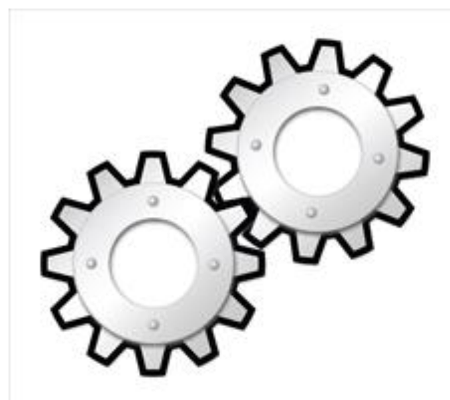
- **Good for:**
  - Summarizing data for presentation
  - Answering questions rigorously
  - Making predictions
  - Making rational, evidence-based decisions
  - A long accumulated experience!

STATISTICAL TOOLS

STATISTICAL TOOLS

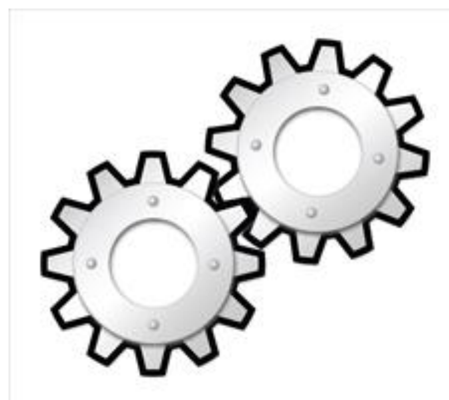
DESCRIPTIVE STATISTICS

INFERENTIAL STATISTICS



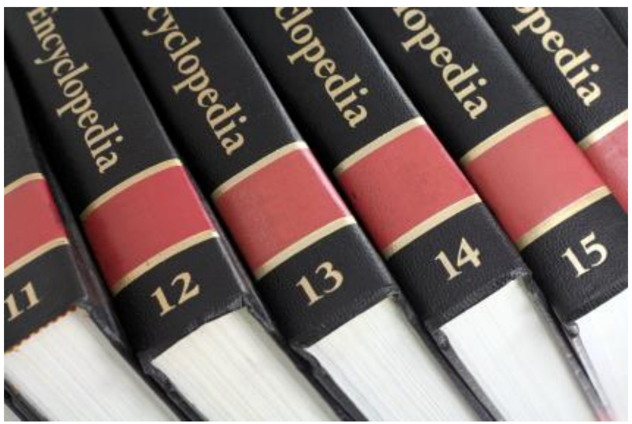
STATISTICAL TOOLS

DESCRIPTIVE STATISTICS



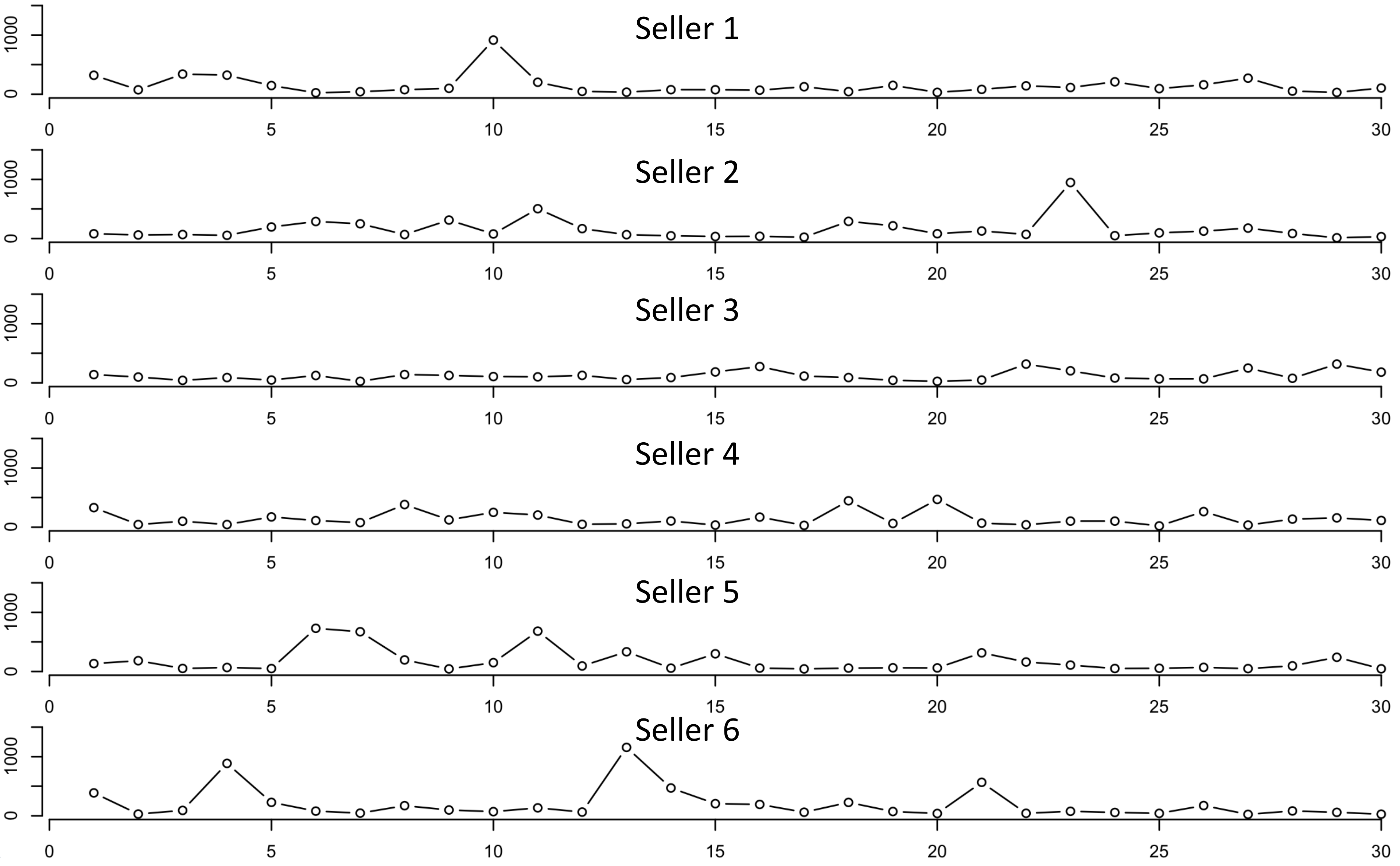
AN EXAMPLE

- Selling encyclopedias

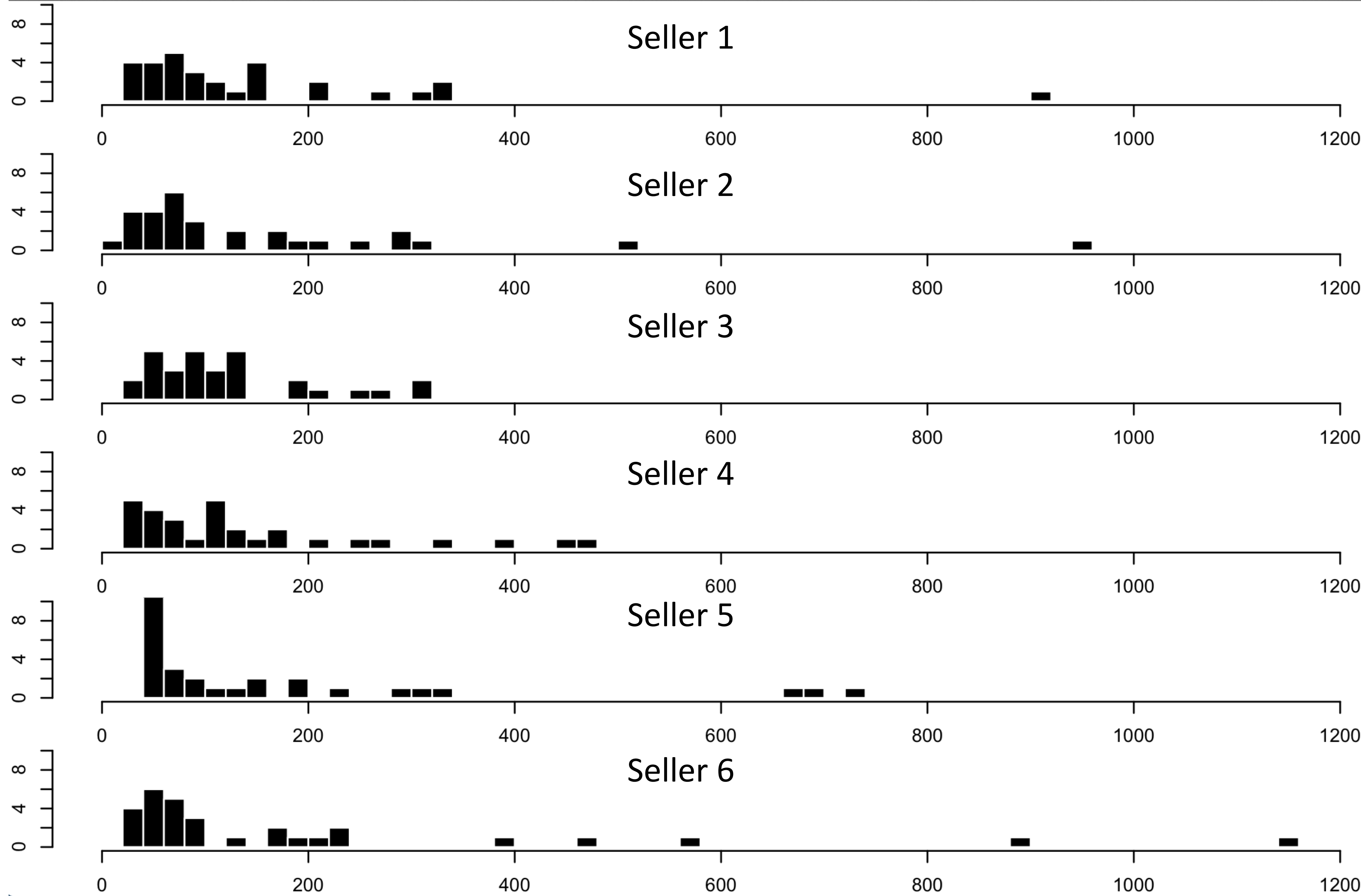


day	Seller 1	Seller 2	Seller 3	Seller 4	Seller 5	Seller 6
1	€320	€80	€139	€330	€133	€387
2	€74	€60	€98	€44	€182	€29
3	€340	€67	€42	€100	€51	€91
4	€322	€54	€89	€44	€67	€886
5	€146	€195	€47	€173	€49	€227
6	€24	€288	€124	€111	€730	€79
7	€42	€249	€26	€77	€672	€45
8	€76	€67	€140	€382	€195	€171
9	€99	€312	€125	€123	€43	€98
10	€915	€77	€106	€250	€149	€70
11	€202	€504	€101	€205	€682	€134
12	€47	€167	€126	€48	€93	€63
13	€34	€65	€55	€56	€333	€1,157
14	€76	€46	€89	€104	€56	€470
15	€75	€34	€184	€35	€299	€205
16	€68	€37	€275	€170	€57	€192

day	Seller 1	Seller 2	Seller 3	Seller 4	Seller 5	Seller 6
1	€320	€80	€139	€330	€133	€387
2	€74	€60	€98	€44	€182	€29
3	€340	€67	€42	€100	€51	€91
4	€322	€54	€89	€44	€67	€886
5	€146	€195	€47	€173	€49	€227
6	€24	€288	€124	€111	€730	€79
7	€42	€249	€26	€77	€672	€45
8	€76	€67	€140	€382	€195	€171
9	€99	€312	€125	€123	€43	€98
10	€915	€77	€106	€250	€149	€70
11	€202	€504	€101	€205	€682	€134
12	€47	€167	€126	€48	€93	€63
13	€34	€65	€55	€56	€333	€1,157
14	€76	€46	€89	€104	€56	€470
15	€75	€34	€184	€35	€299	€205
16	€68	€37	€275	€170	€57	€192
17	€126	€23	€114	€30	€43	€60
18	€43	€290	€89	€446	€57	€226
19	€149	€215	€43	€63	€62	€72
20	€31	€81	€26	€469	€60	€39
21	€81	€127	€47	€68	€315	€566
22	€141	€70	€317	€40	€160	€42
23	€113	€947	€203	€102	€108	€76
24	€209	€48	€81	€102	€50	€56
25	€94	€95	€67	€21	€54	€41
26	€159	€125	€67	€263	€69	€173
27	€271	€176	€250	€35	€48	€24
28	€52	€85	€77	€136	€95	€82
29	€30	€12	€317	€157	€240	€58
30	€104	€31	€181	€113	€45	€27



Seller 1
Seller 2
Seller 3
Seller 4
Seller 5
Seller 6
0
1000
0
5
10
15
20
25
30

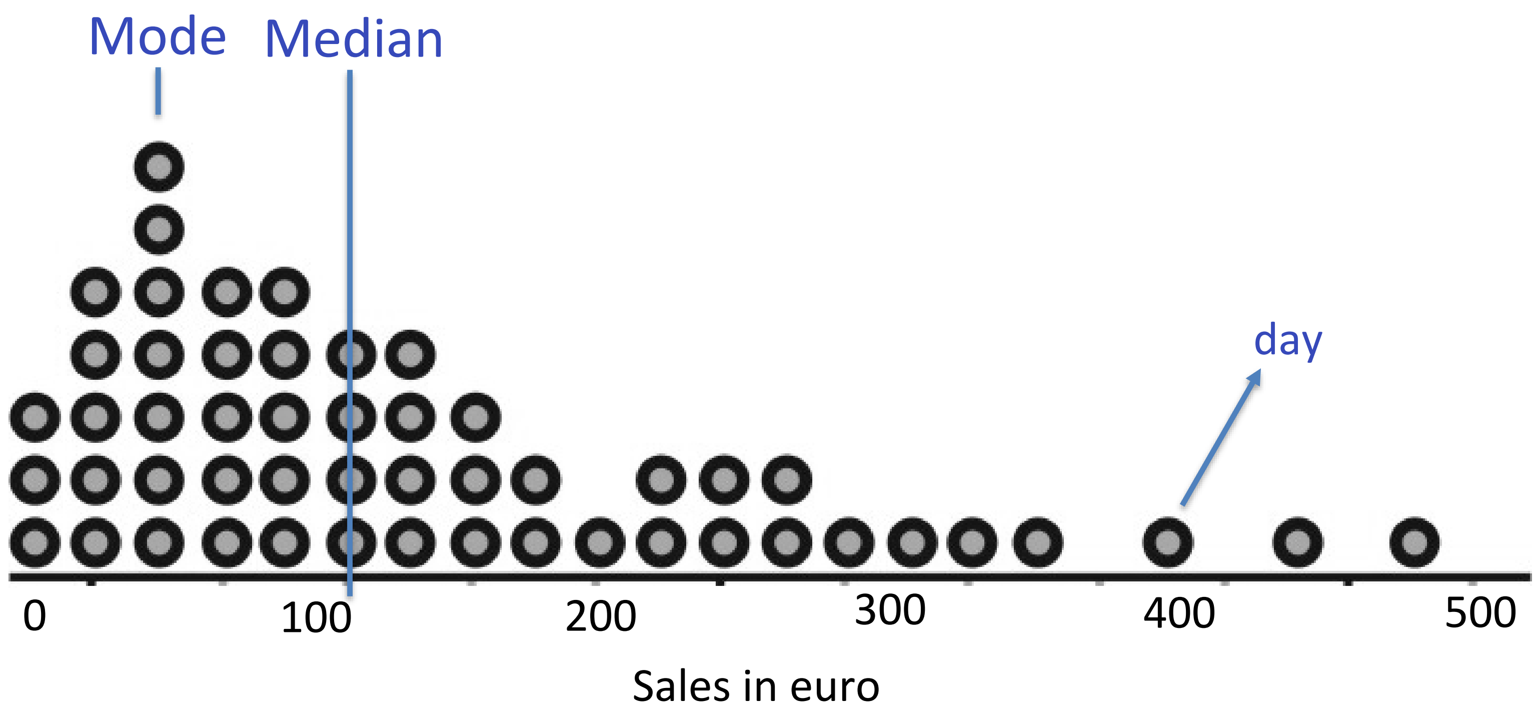

Seller 1
Seller 2
Seller 3
Seller 4
Seller 5
Seller 6
0
4
8
0
200
400
600
800
1000
1200

CENTRAL TENDENCY

Name & Meaning	Formula / Example	Used for
Arithmetic Mean [average]	$\frac{sum}{size} = \frac{a+b+c}{3}$	Most situations ("average item")
Median [middle value]	Middle of sorted list (2 middles? Average 'em)	Wildly varying samples (houses, incomes)
Mode [most popular]	Most popular value	No compromises (winner takes all)
Geometric Mean [average factor]	$\sqrt[3]{abc}$	Investments, growth, area, volume
Harmonic Mean [average rate]	$\frac{3}{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}}$	Speed, production, cost

From Kalid Azad

CENTRAL TENDENCY

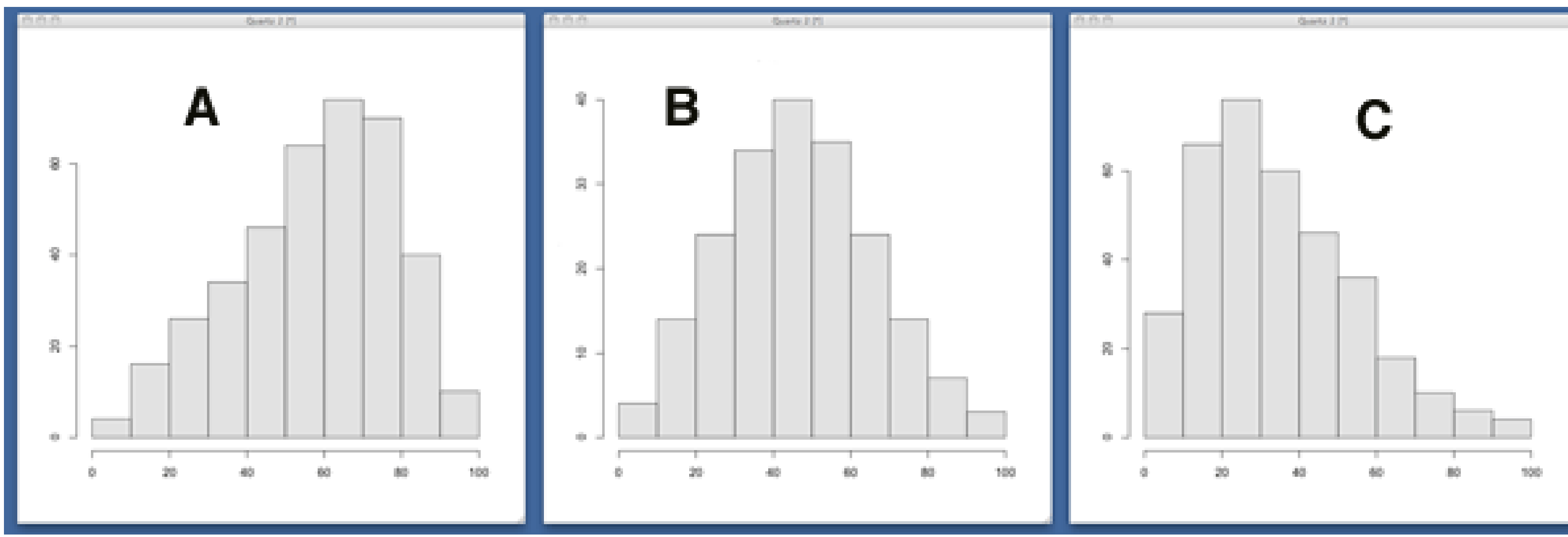


CENTRAL TENDENCY

When are the mean and the median equal? When do they differ?

CENTRAL TENDENCY

negative skew | symmetric | positive skew

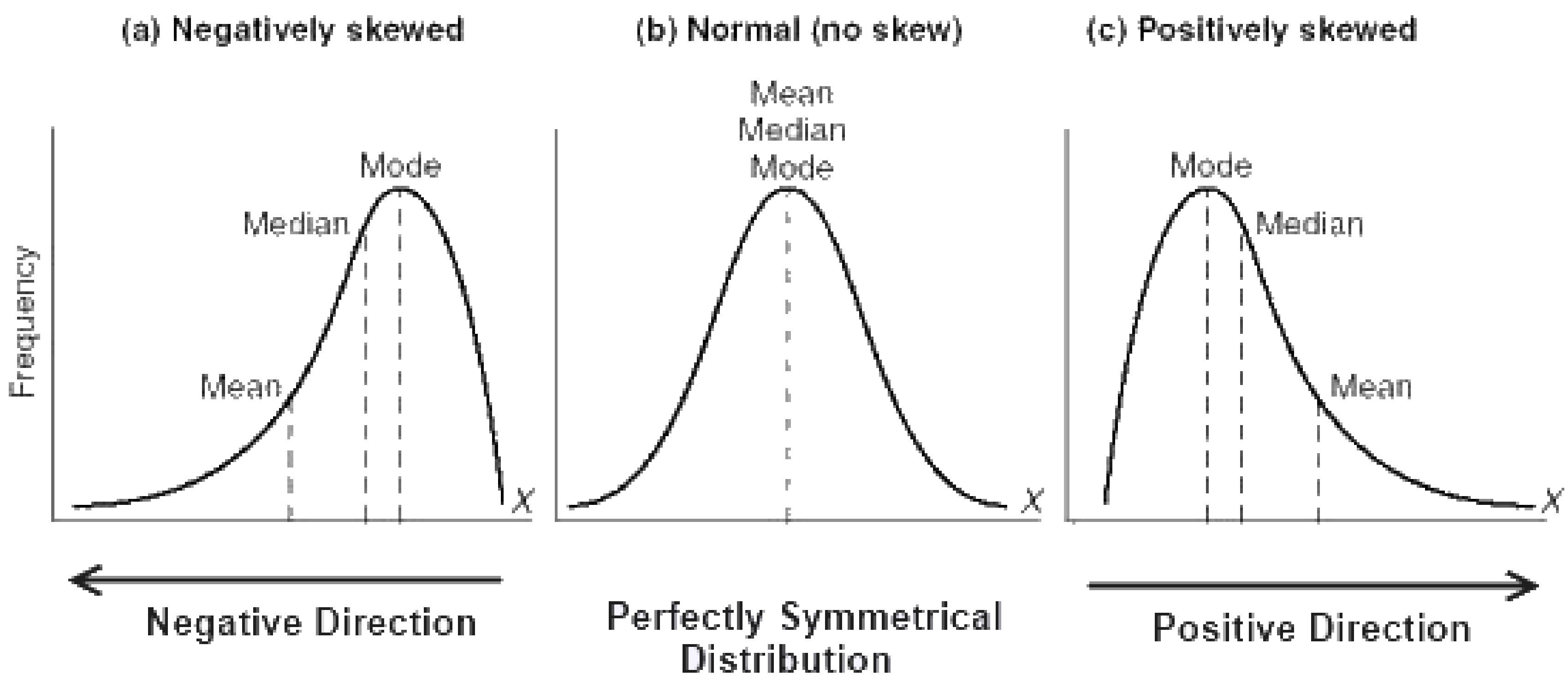


From Shreya Sethi

CENTRAL TENDENCY

CENTRAL TENDENCY

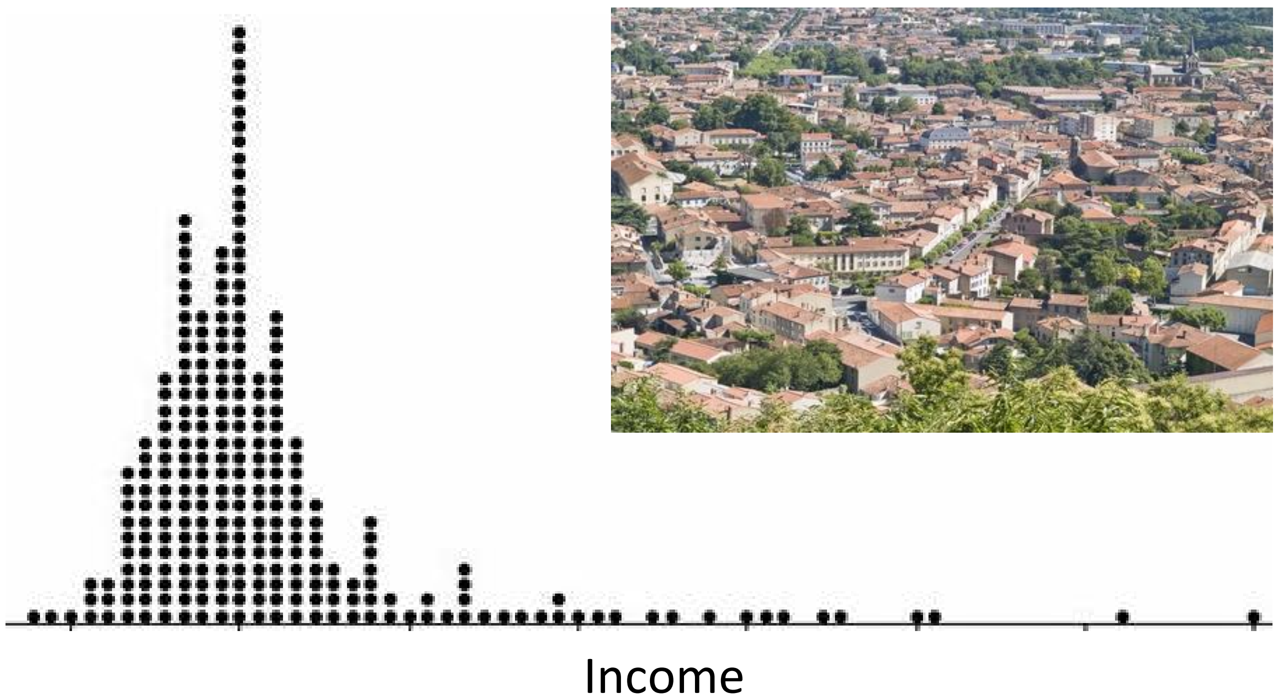
CENTRAL TENDENCY



From Shreya Sethi

CENTRAL TENDENCY

What is the best measure of central tendency?



DISPERSION

Standard Deviation

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

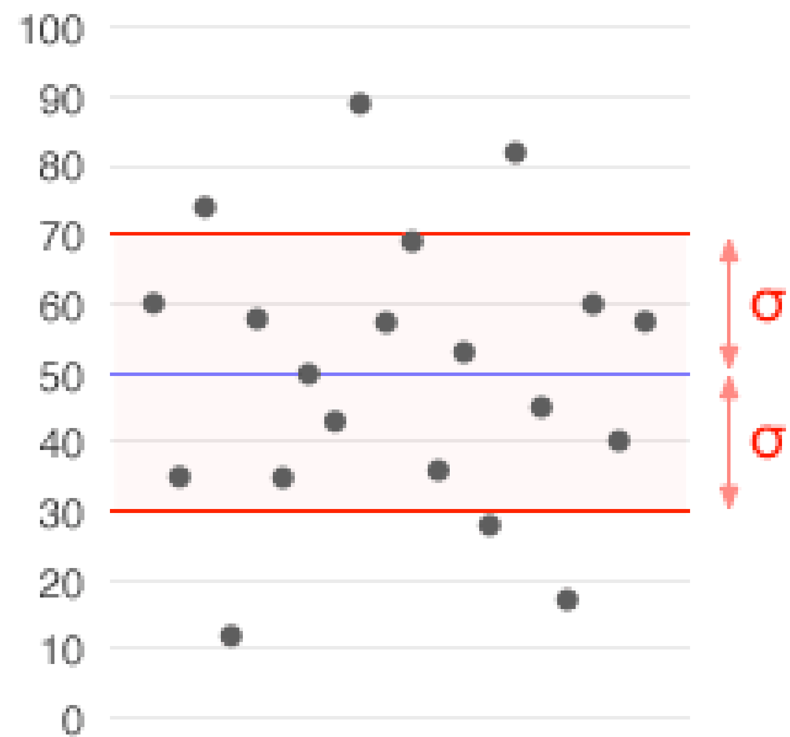
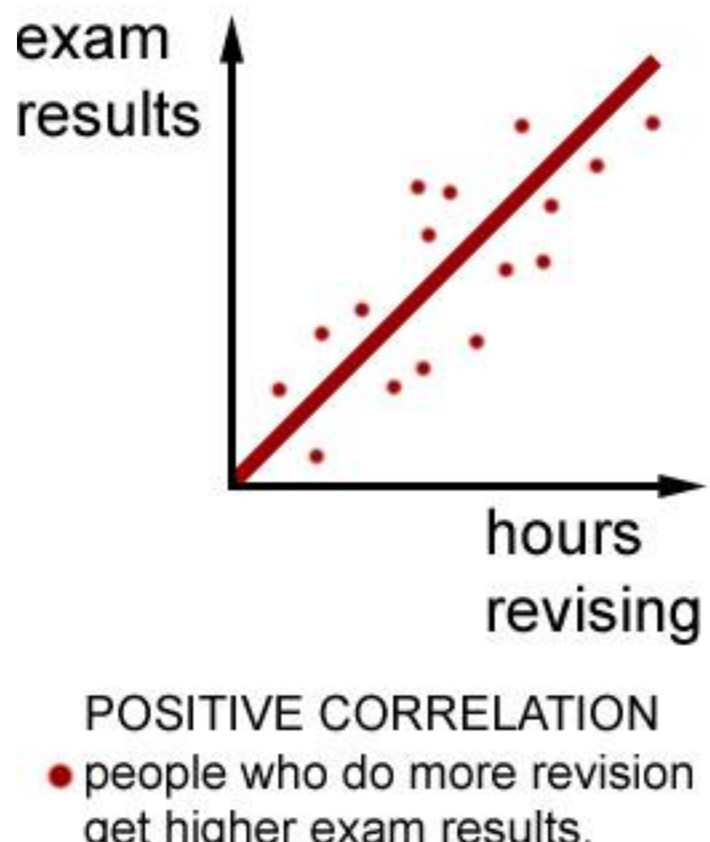


Image from Wikipedia

DEPENDENCE

Correlation



Source unknown

DEPENDENCE

- Correlation

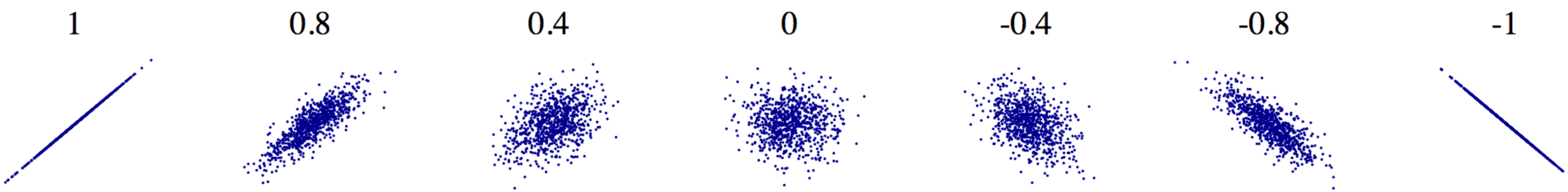
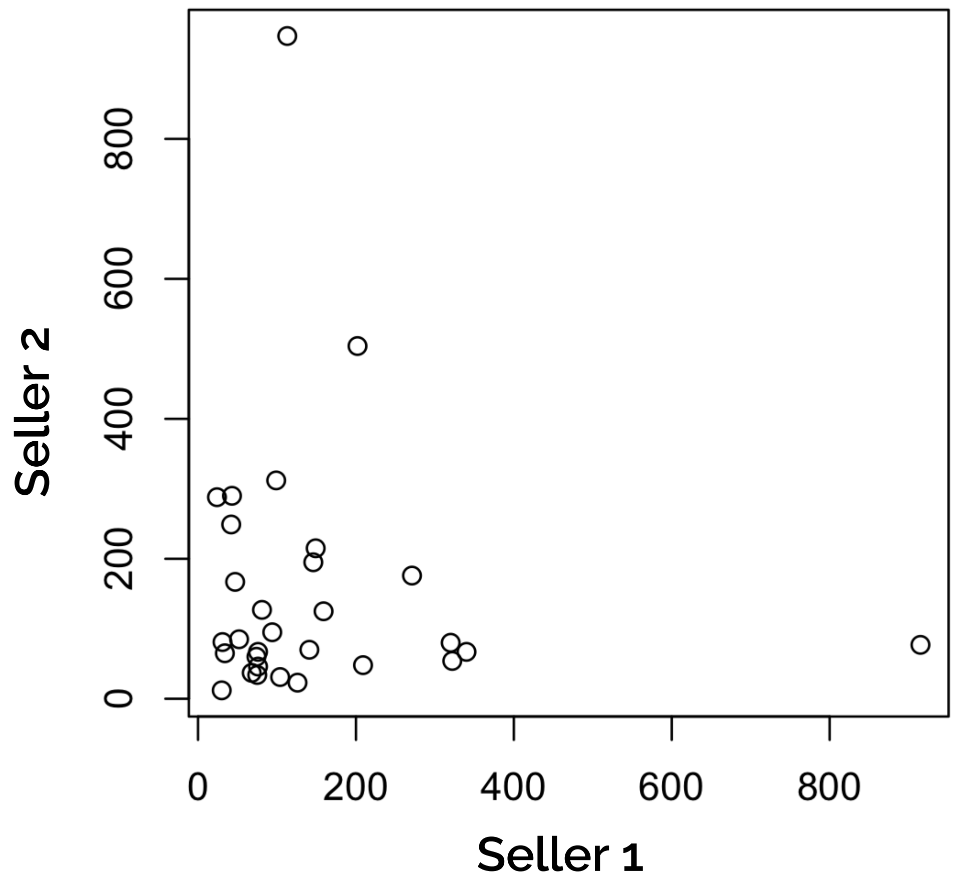


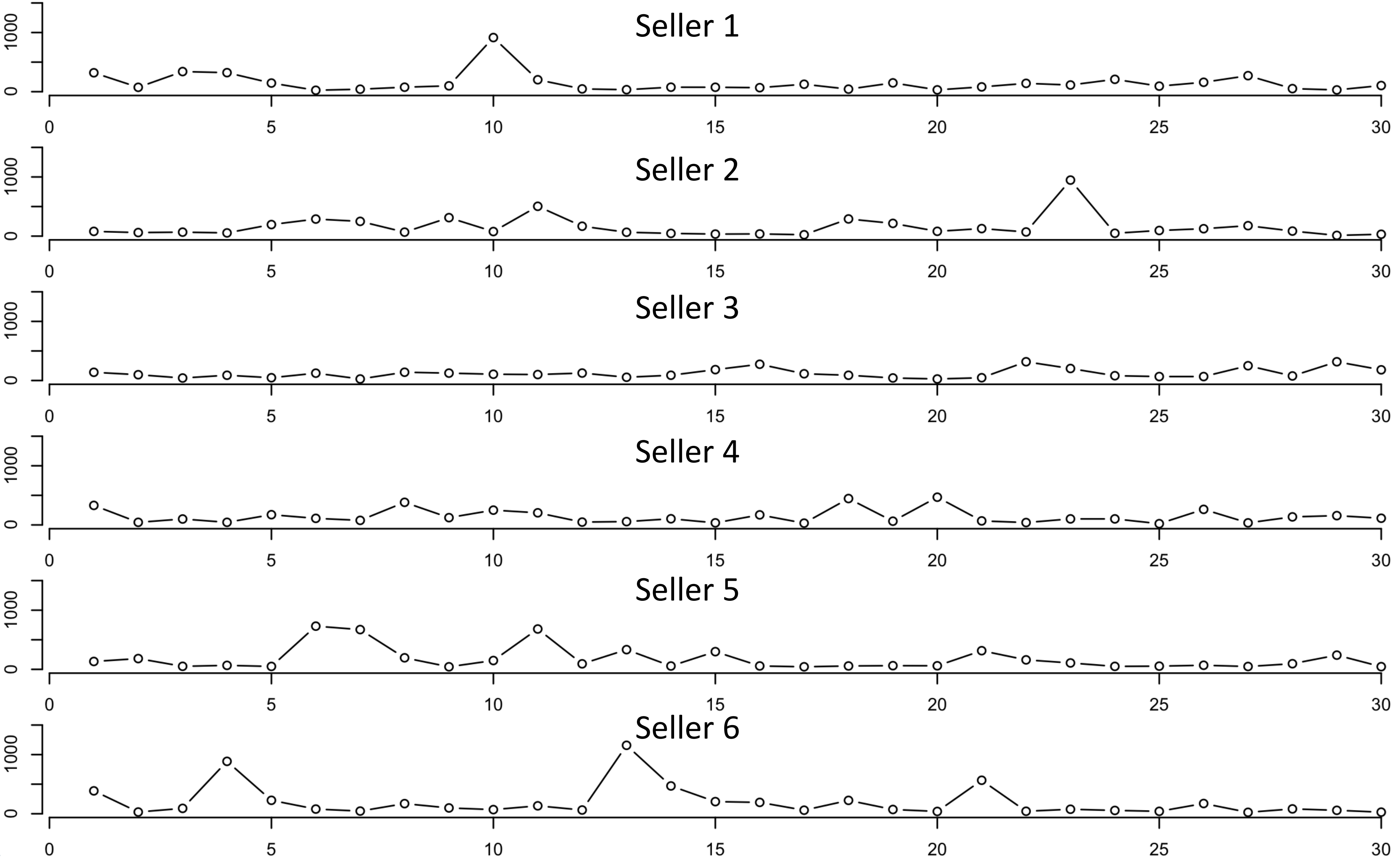
Image from Wikipedia

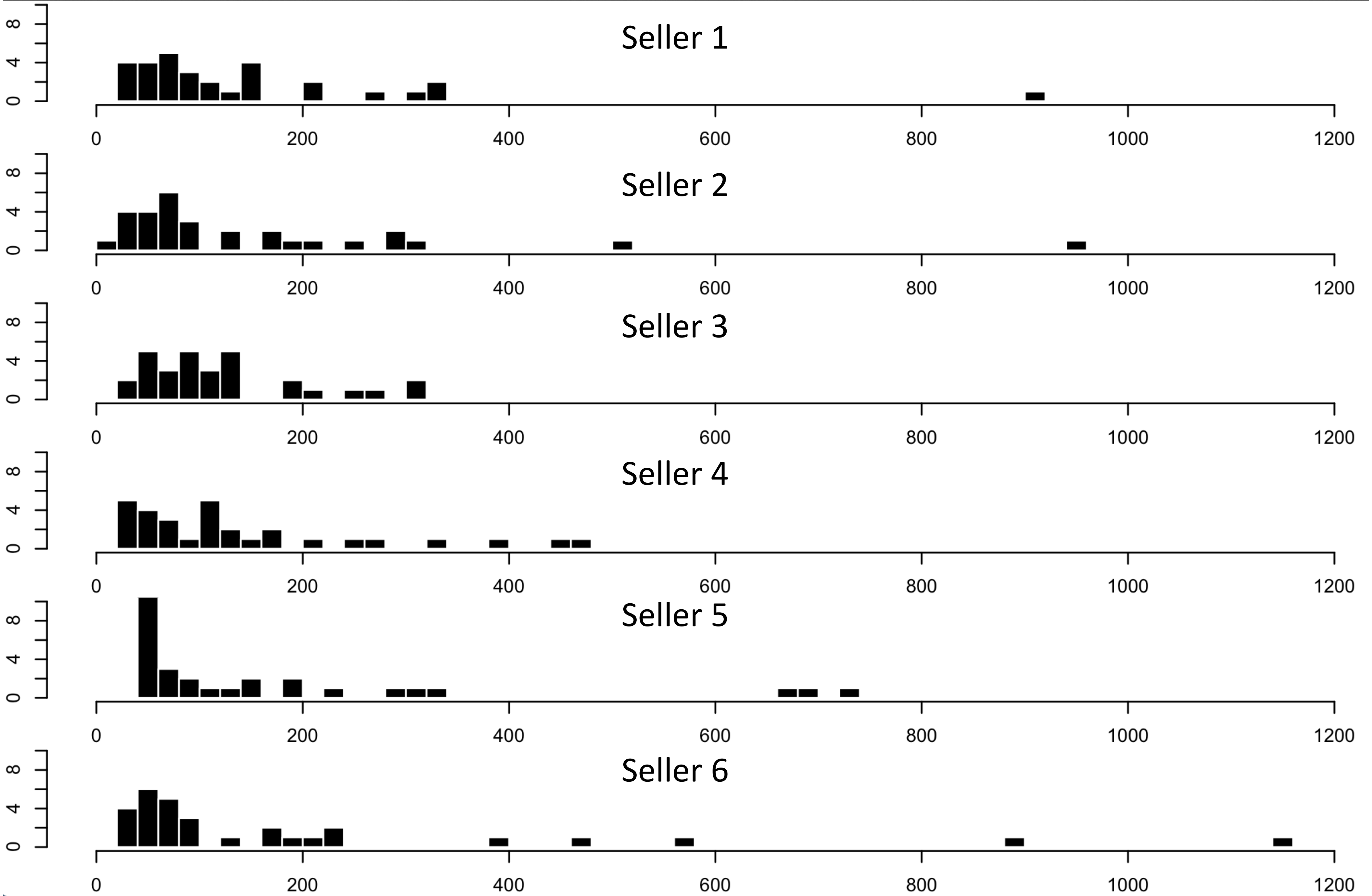
DEPENDENCE

Correlation

r = -0.08




Seller 1
Seller 2
Seller 3
Seller 4
Seller 5
Seller 6
0
1000
0
5
10
15
20
25
30

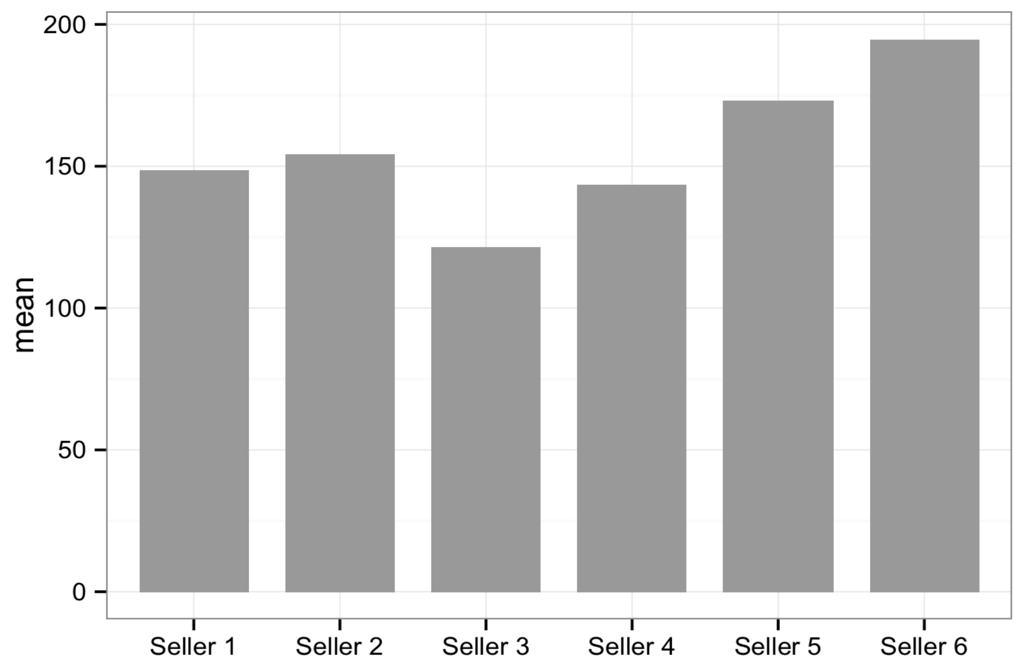

Seller 1
Seller 2
Seller 3
Seller 4
Seller 5
Seller 6
0
4
8
0
200
400
600
800
1000
1200

Average Sales

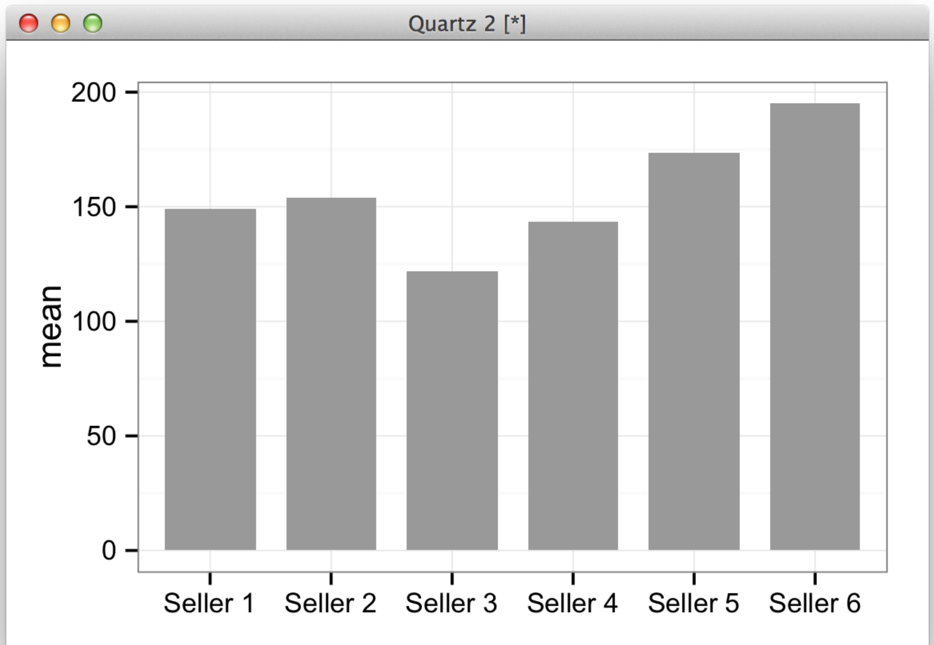
Seller 1	Seller 2	Seller 3	Seller 4	Seller 5	Seller 6
€149	€154	€122	€143	€173	€195

Average Sales

Seller 1	Seller 2	Seller 3	Seller 4	Seller 5	Seller 6
€149	€154	€122	€143	€173	€195



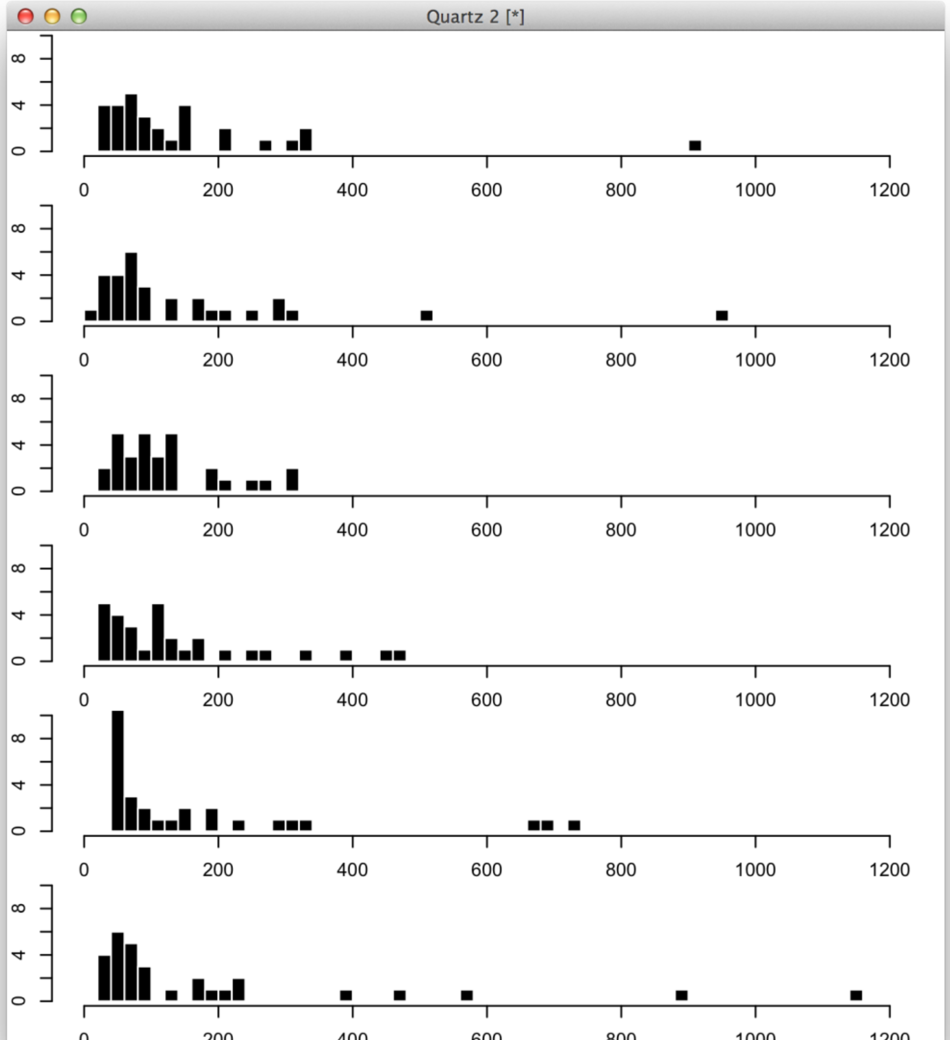
September 2014



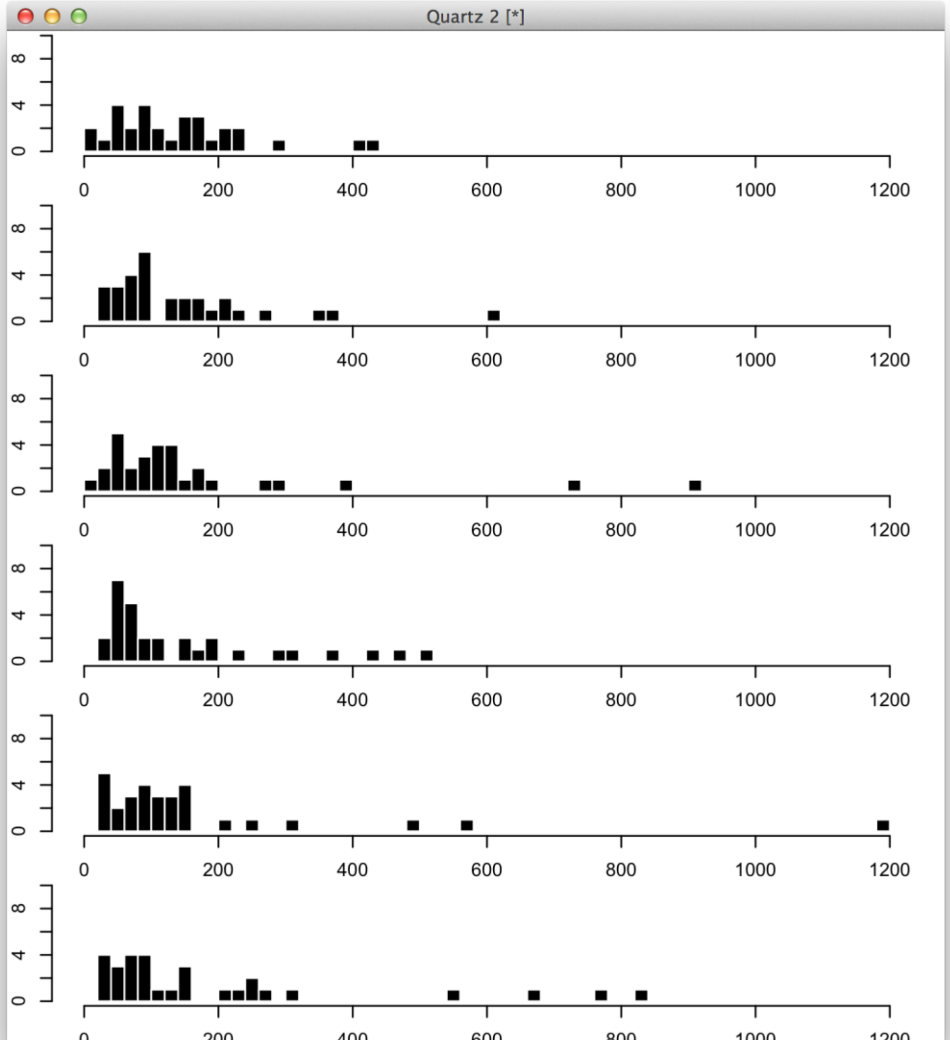
How much can we trust this chart?

LET US TRAVEL TO THE FUTURE

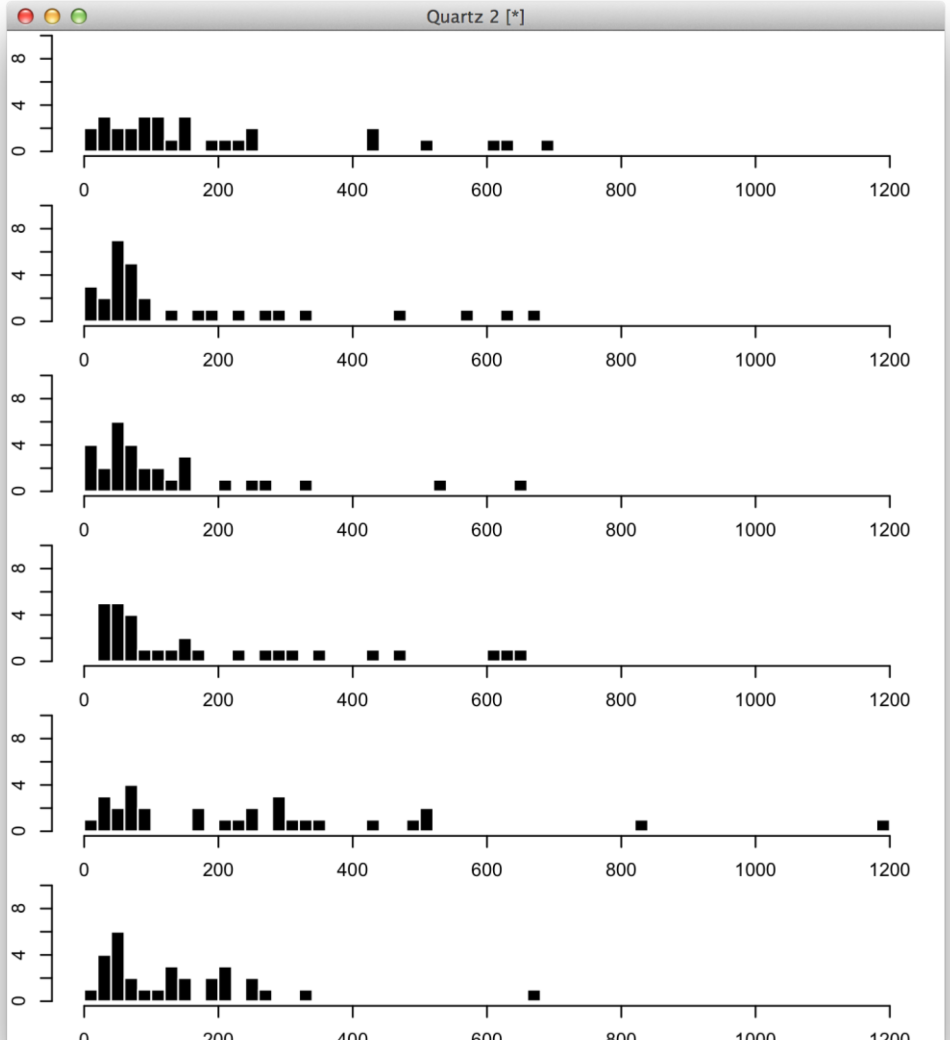
September 2014



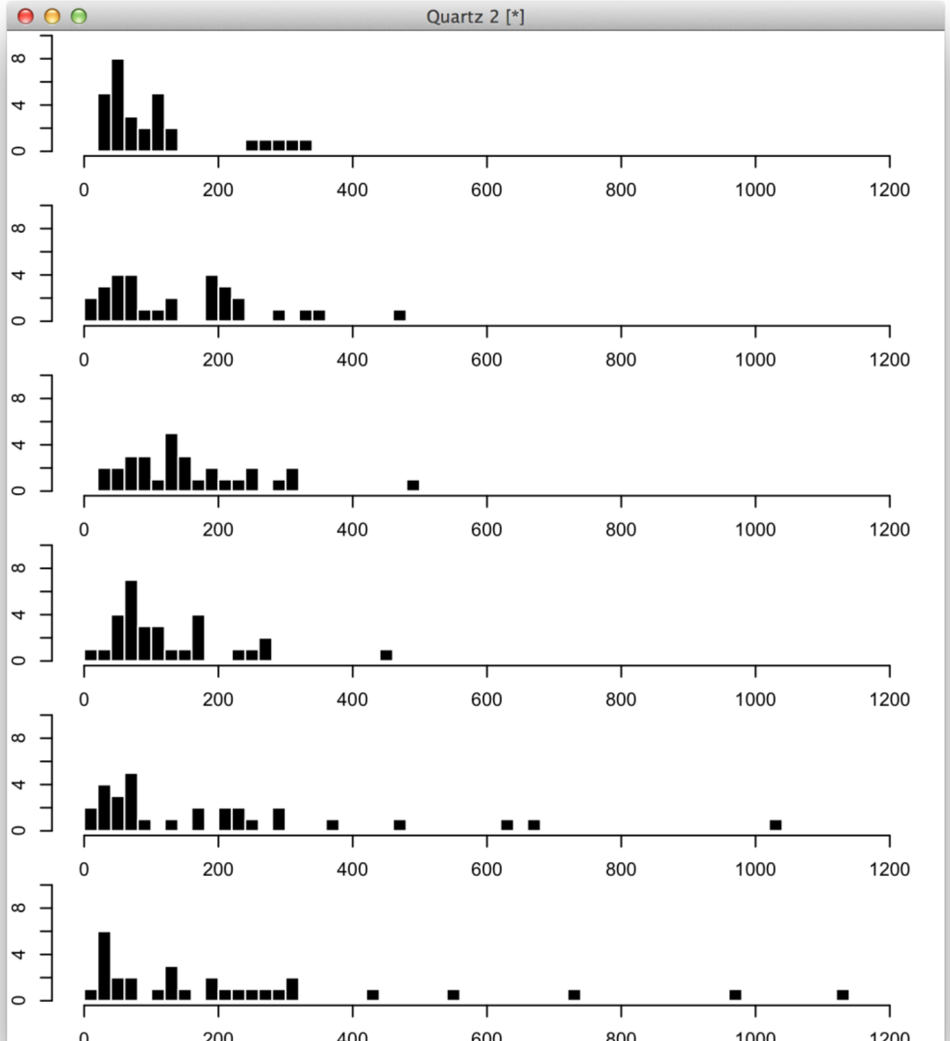
October 2014



November 2014



December 2014



September 2014

October 2014

November 2014

December 2014

BACK TO THE PRESENT

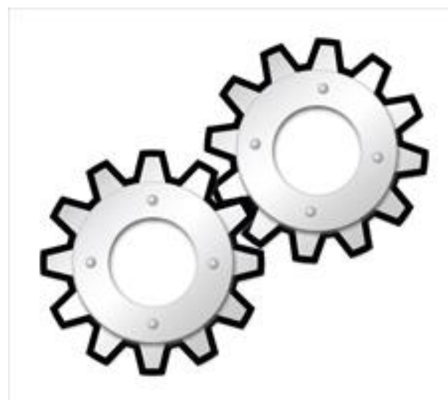
September 2014

day	Seller 1	Seller 2	Seller 3	Seller 4	Seller 5	Seller 6
1	€320	€80	€139	€330	€133	€387
2	€74	€60	€98	€44	€182	€29
3	€340	€67	€42	€100	€51	€91
4	€322	€54	€89	€44	€67	€886
5	€146	€195	€47	€173	€49	€227
6	€24	€288	€124	€111	€730	€79
7	€42	€249	€26	€77	€672	€45
8	€76	€67	€140	€382	€195	€171
9	€99	€312	€125	€123	€43	€98
10	€915	€77	€106	€250	€149	€70
11	€202	€504	€101	€205	€682	€134
12	€47	€167	€126	€48	€93	€63
13	€34	€65	€55	€56	€333	€1,157
14	€76	€46	€89	€104	€56	€470
15	€75	€34	€184	€35	€299	€205
16	€68	€37	€275	€170	€57	€192

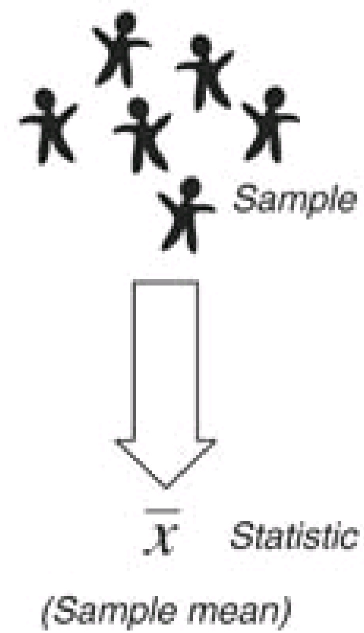
How much can we trust this chart?

STATISTICAL TOOLS

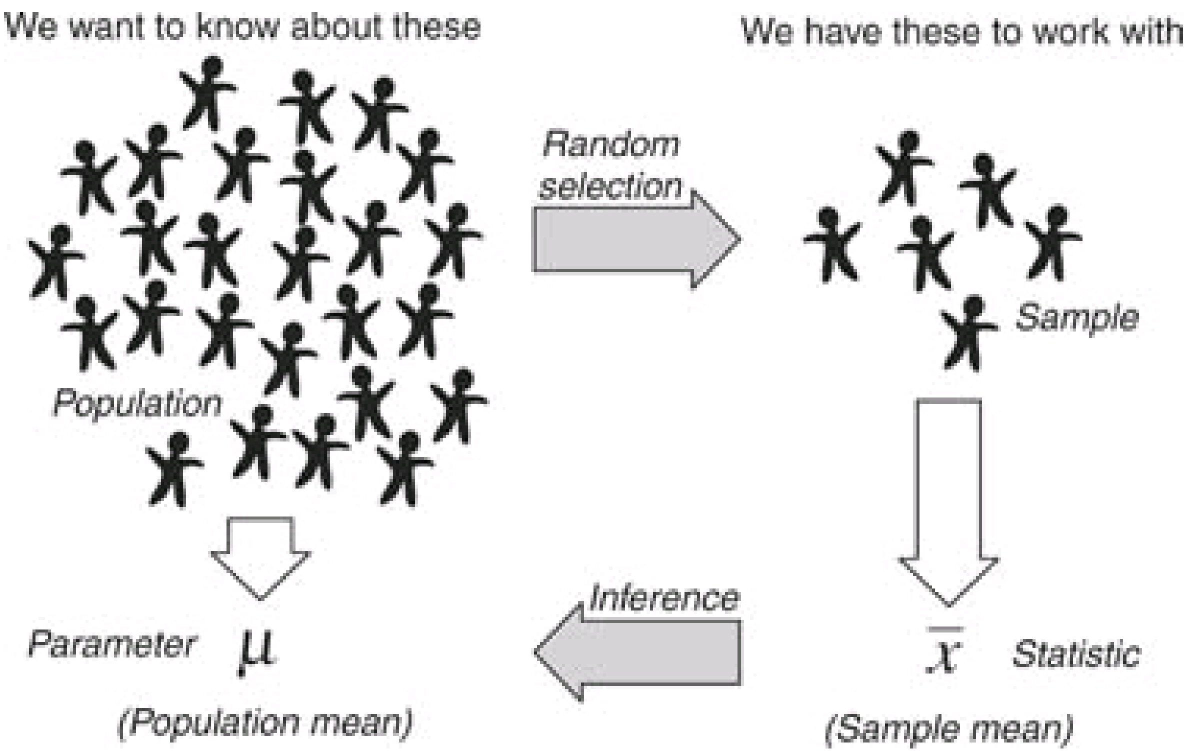
INFERENTIAL STATISTICS



STATISTICAL INFERENCE



STATISTICAL INFERENCE



STATISTICAL INFERENCE

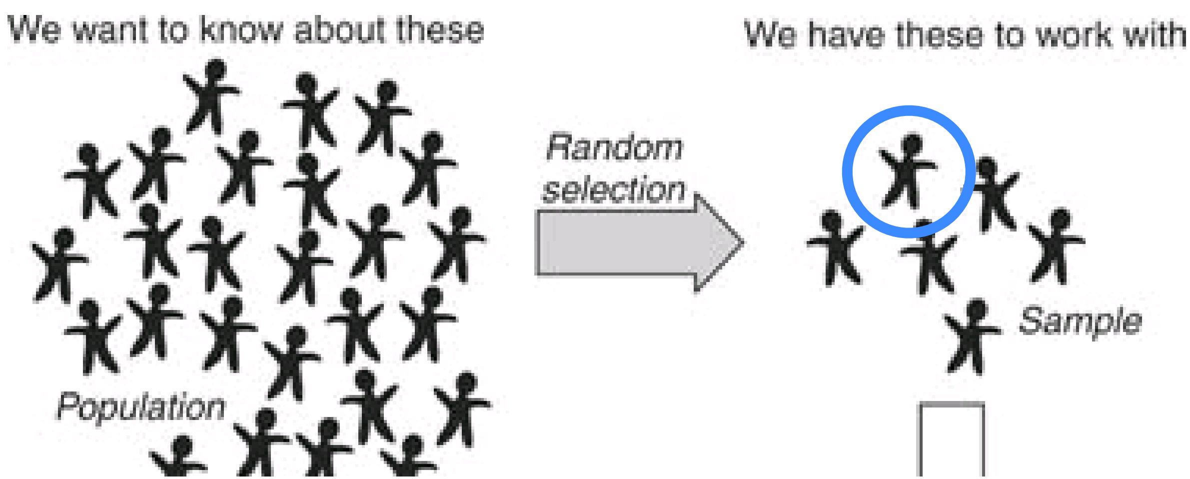
Terminology:

Sample vs. population

- Mean, median, standard deviation, correlation, etc:
  - A sample statistic
  - A population parameter

STATISTICAL INFERENCE

Unit of statistical analysis



= "the thing that I'm sampling from a larger population"

STATISTICAL INFERENCE

Unit of statistical analysis

day	Seller 1	Seller 2	Seller 3	Seller 4	Seller 5	Seller 6
1	€320	€80	€139	€330	€133	€387
2	€74	€60	€98	€44	€182	€29
3	€340	€67	€42	€100	€51	€91
4	€322	€54	€89	€44	€67	€886
5	€146	€195	€47	€173	€49	€227
6	€24	€288	€124	€111	€730	€79
7	€42	€249	€26	€77	€672	€45
8	€76	€67	€140	€382	€195	€171
9	€99	€312	€125	€123	€43	€98
10	€915	€77	€106	€250	€149	€70
11	€202	€504	€101	€205	€682	€134

STATISTICAL INFERENCE

Unit of statistical analysis

day	Seller 1
1	€320
2	€74
3	€340
4	€322
5	€146
6	€24
7	€42
8	€76
9	€99
10	€915

STATISTICAL INFERENCE

Unit of statistical analysis

day	Seller 1	Seller 2	Seller 3	Seller 4	Seller 5	Seller 6
1	€320	€80	€139	€330	€133	€387
2	€74	€60	€98	€44	€182	€29
3	€340	€67	€42	€100	€51	€91
4	€322	€54	€89	€44	€67	€886
5	€146	€195	€47	€173	€49	€227
6	€24	€288	€124	€111	€730	€79
7	€42	€249	€26	€77	€672	€45
8	€76	€67	€140	€382	€195	€171
9	€99	€312	€125	€123	€43	€98
10	€915	€77	€106	€250	€149	€70
11	€202	€504	€101	€205	€682	€134

STATISTICAL INFERENCE

Unit of statistical analysis

Average Sales

Seller 1	Seller 2	Seller 3	Seller 4	Seller 5	Seller 6
€149	€154	€122	€143	€173	€195

STATISTICAL INFERENCE

Unit of statistical analysis

day	Seller 1	Seller 2	Seller 3	Seller 4	Seller 5	Seller 6
1	€320	€80	€139	€330	€133	€387
2	€74	€60	€98	€44	€182	€29
3	€340	€67	€42	€100	€51	€91
4	€322	€54	€89	€44	€67	€886
5	€146	€195	€47	€173	€49	€227
6	€24	€288	€124	€111	€730	€79
7	€42	€249	€26	€77	€672	€45
8	€76	€67	€140	€382	€195	€171
9	€99	€312	€125	€123	€43	€98
10	€915	€77	€106	€250	€149	€70
11	€202	€504	€101	€205	€682	€134

SAMPLING DISTRIBUTION

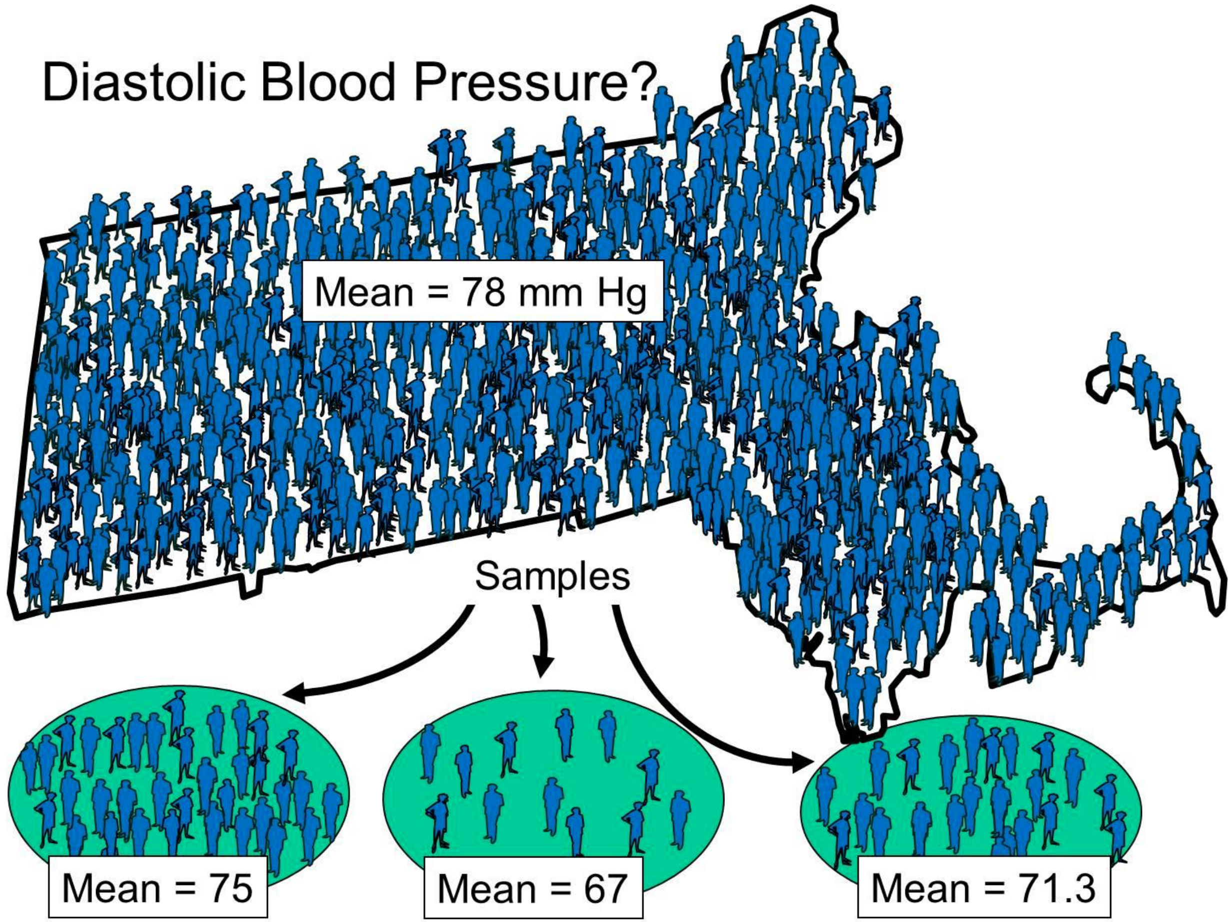
SAMPLING DISTRIBUTION

"The sampling distribution of a statistic is the distribution of that statistic, considered as a random variable, when derived from a random sample of size n."

[...]

"It may be considered as the distribution of the statistic for all possible samples from the same population of a given size"

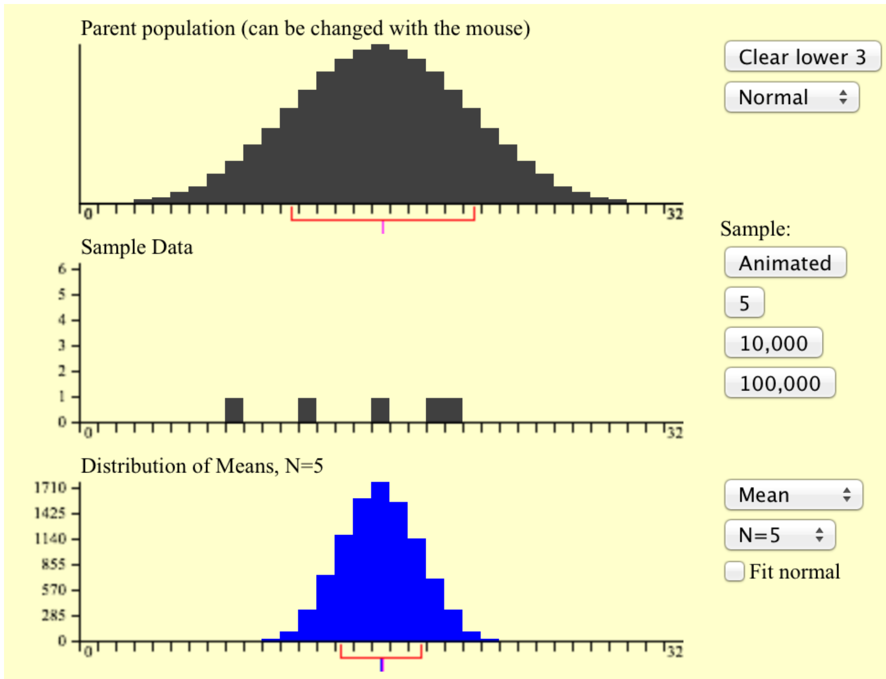
SAMPLING DISTRIBUTION



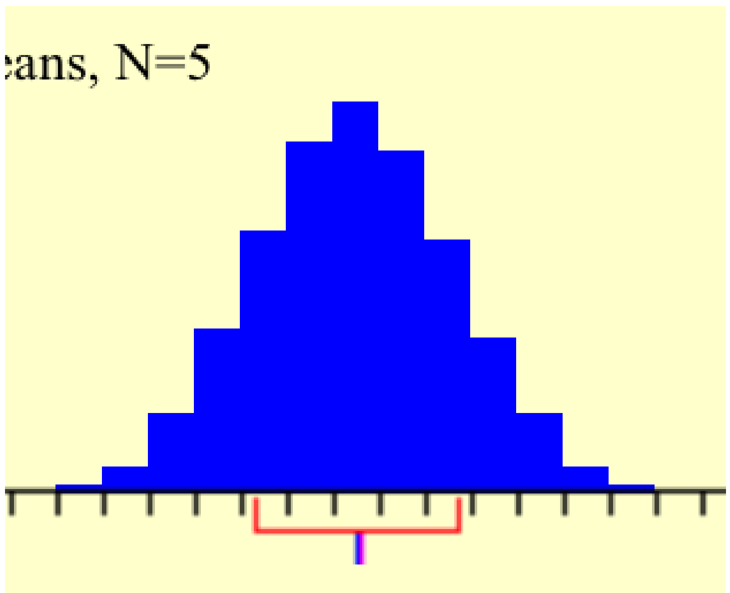
From Lisa Sullivan

SAMPLING DISTRIBUTION

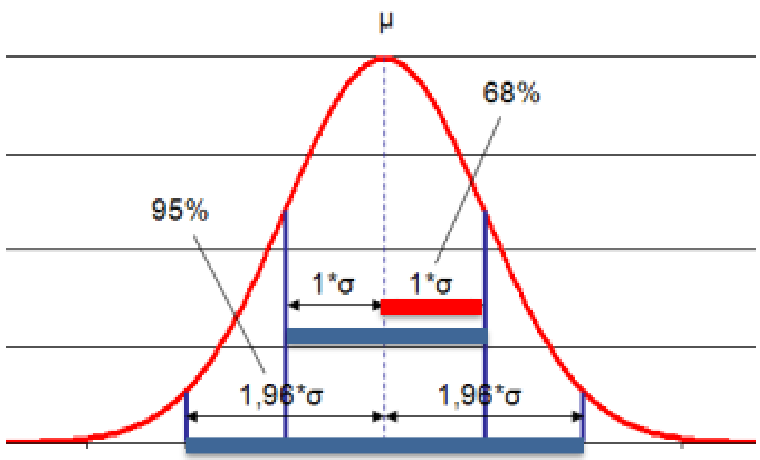
- Demo http://onlinestatbook.com/stat_sim/sampling_dist/



SAMPLING DISTRIBUTION



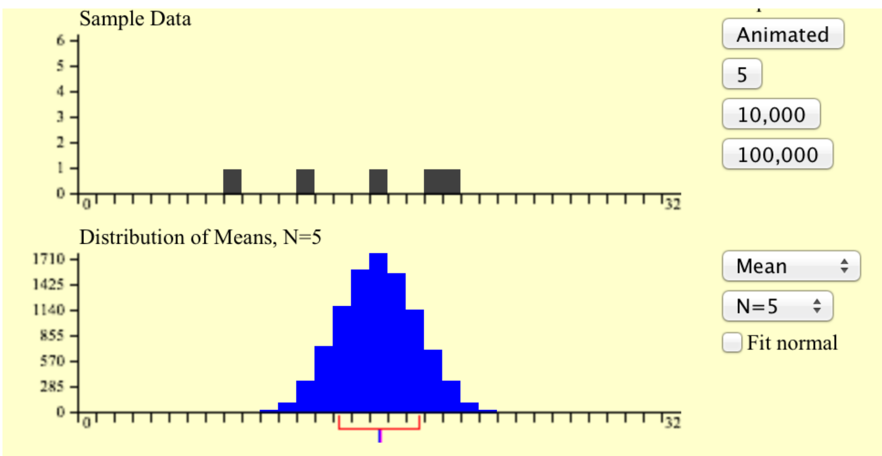
SAMPLING DISTRIBUTION



Standard error

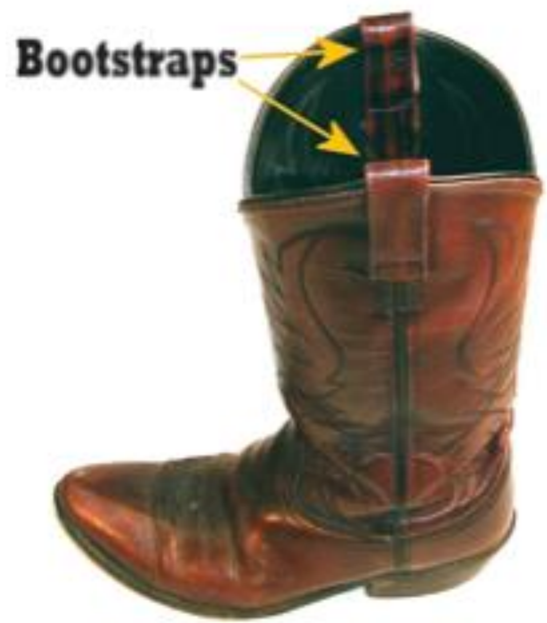
95% confidence interval

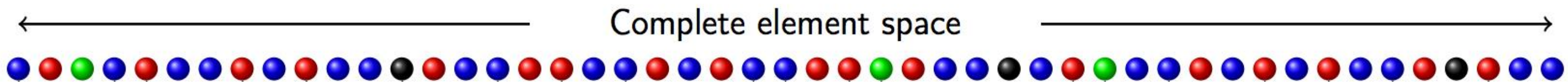
SAMPLING DISTRIBUTION

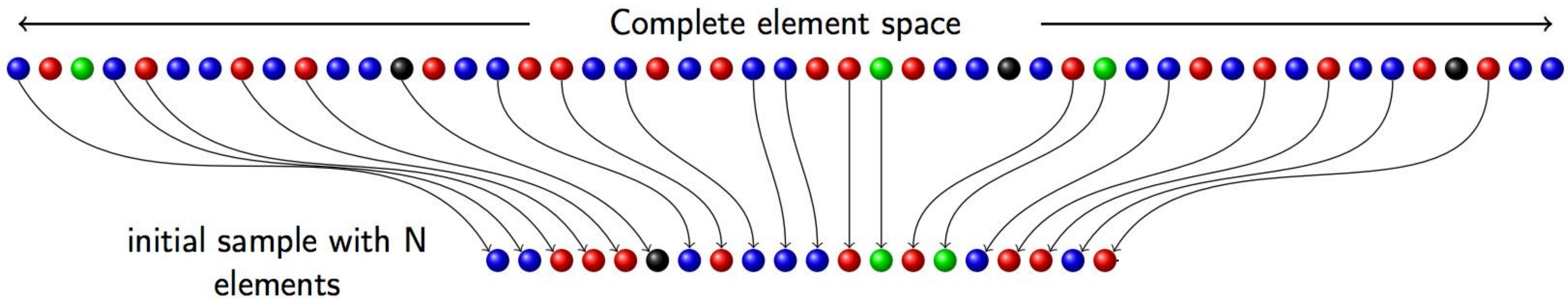


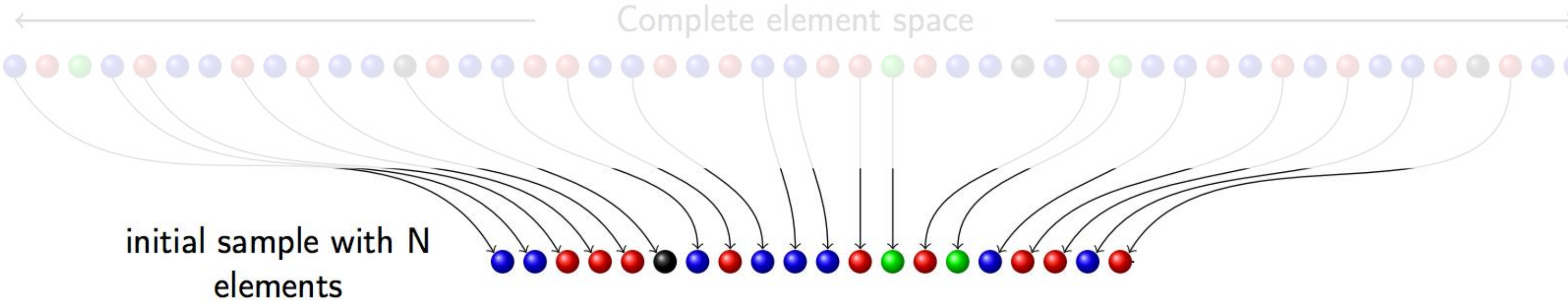
SAMPLING DISTRIBUTION

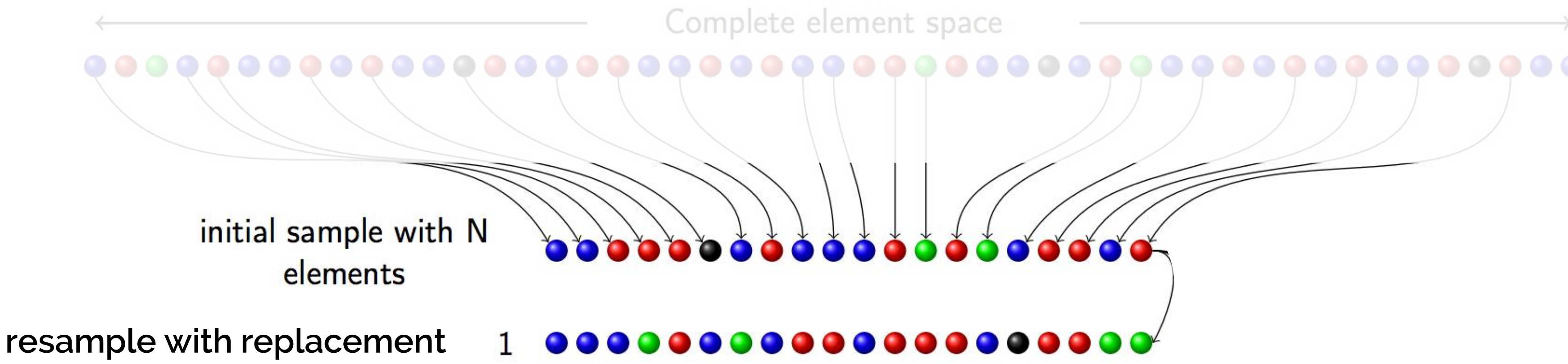
- Resampling techniques
    - Bootstrapping



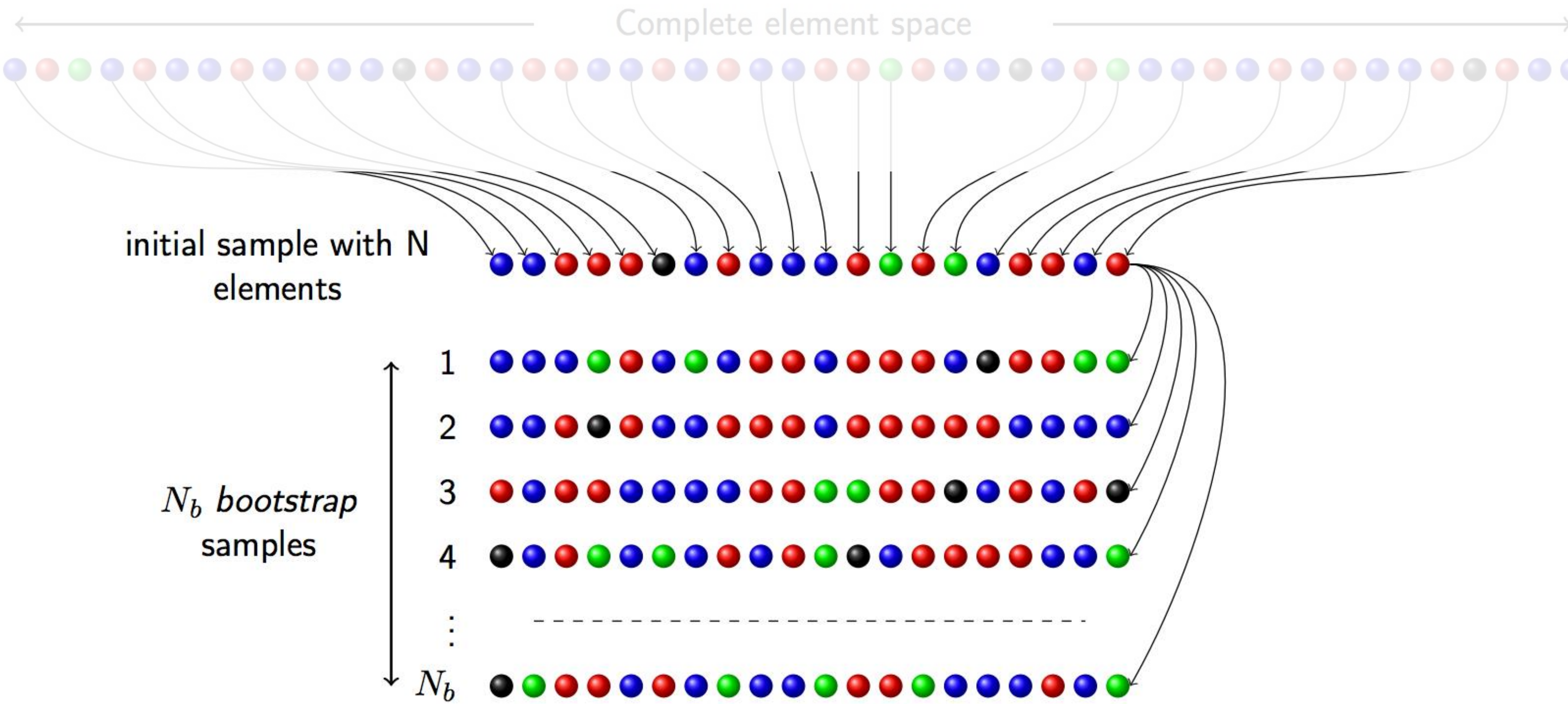

Complete element space

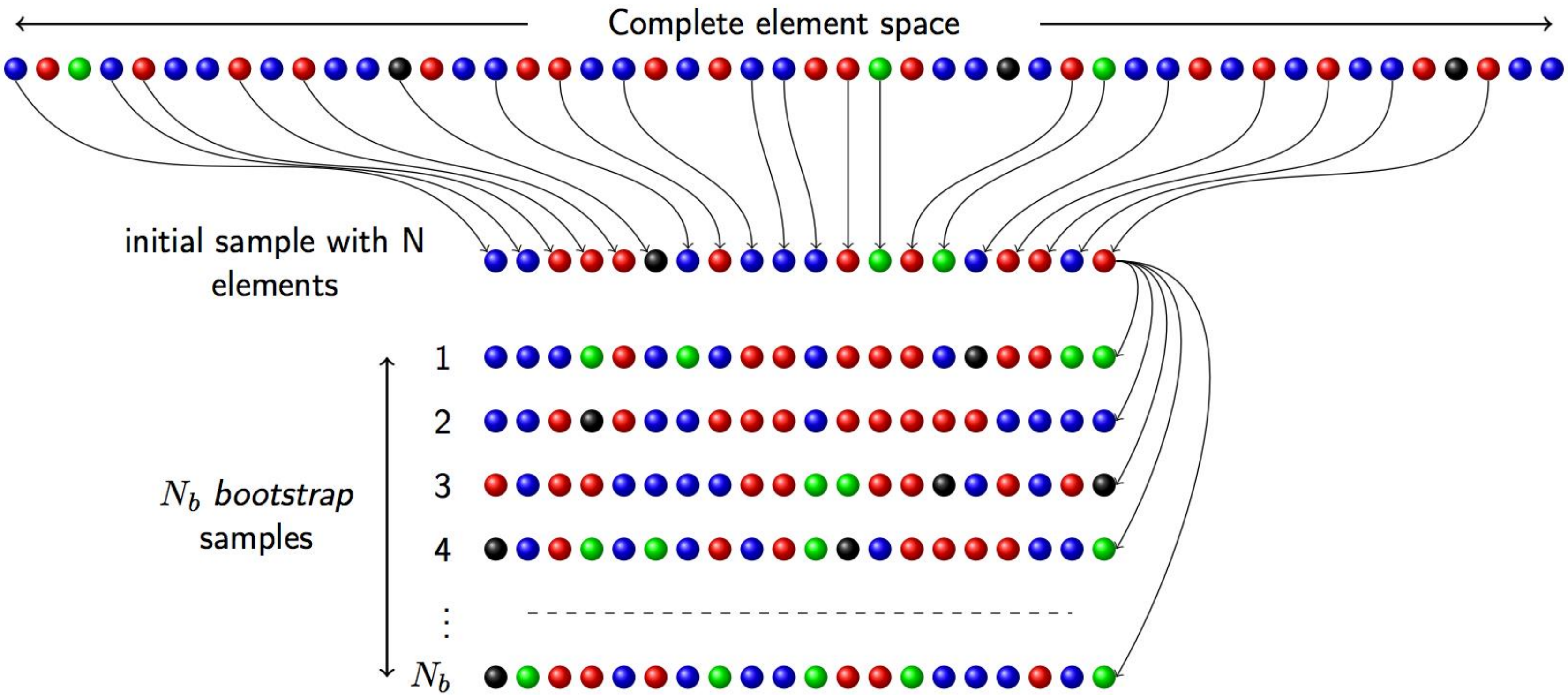

Complete element space
initial sample with N elements


Complete element space
initial sample with N elements



From Germain Salvato-Vallverdu


Complete element space
initial sample with N elements
N_b *bootstrap* samples
1
2
3
4
⋮
N_b

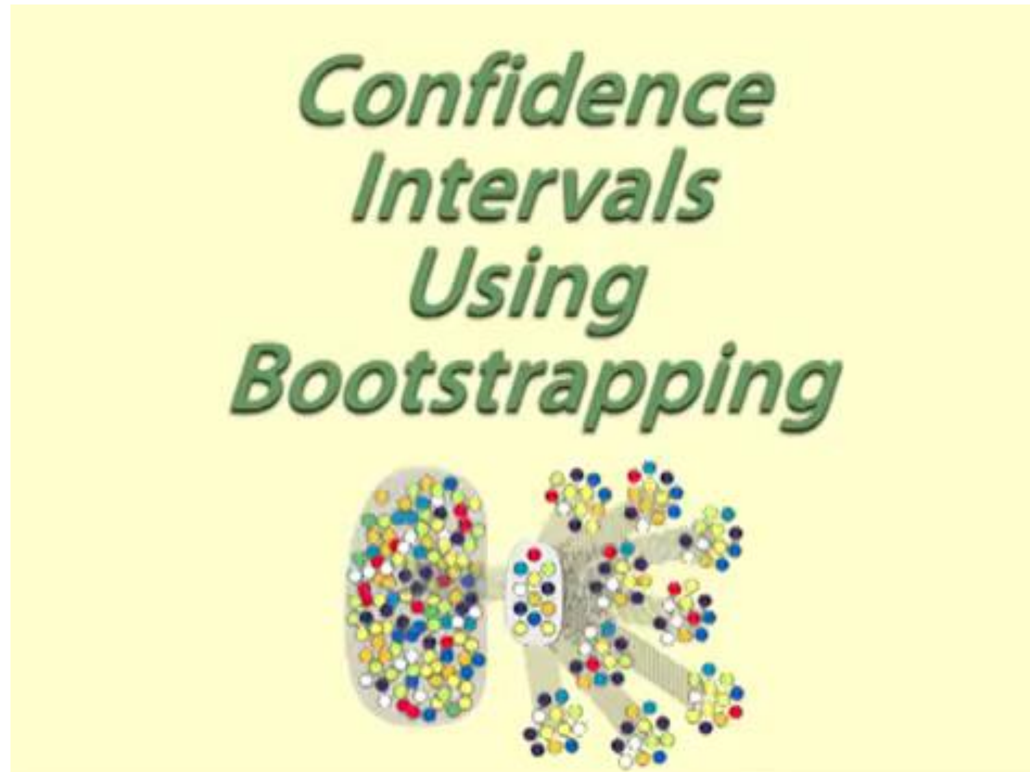


Theorem (B. Efron, Ann. Statist. 1979)

When N tend to infinity, the distribution of average values computed from bootstrap samples is equal to the distribution of average values obtained from ALL samples with N elements which can be constructed from the complete space. Thus the width of the distribution gives an evaluation of the sample quality.

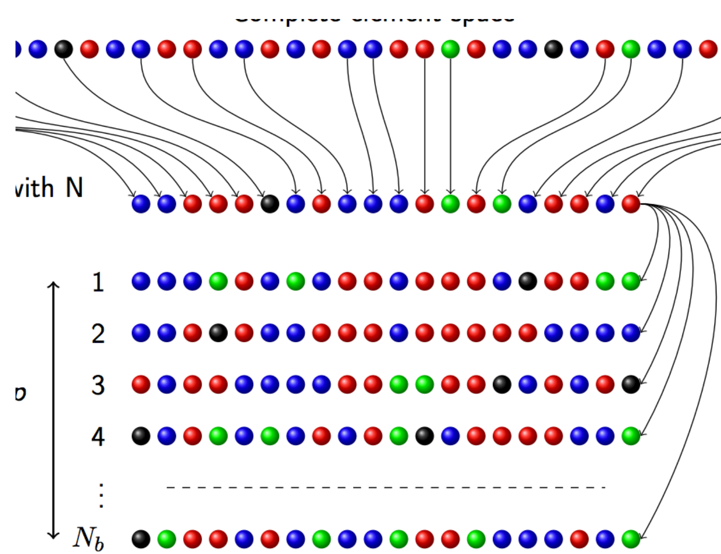
SAMPLING DISTRIBUTION

- Bootstrapping video



SAMPLING DISTRIBUTION

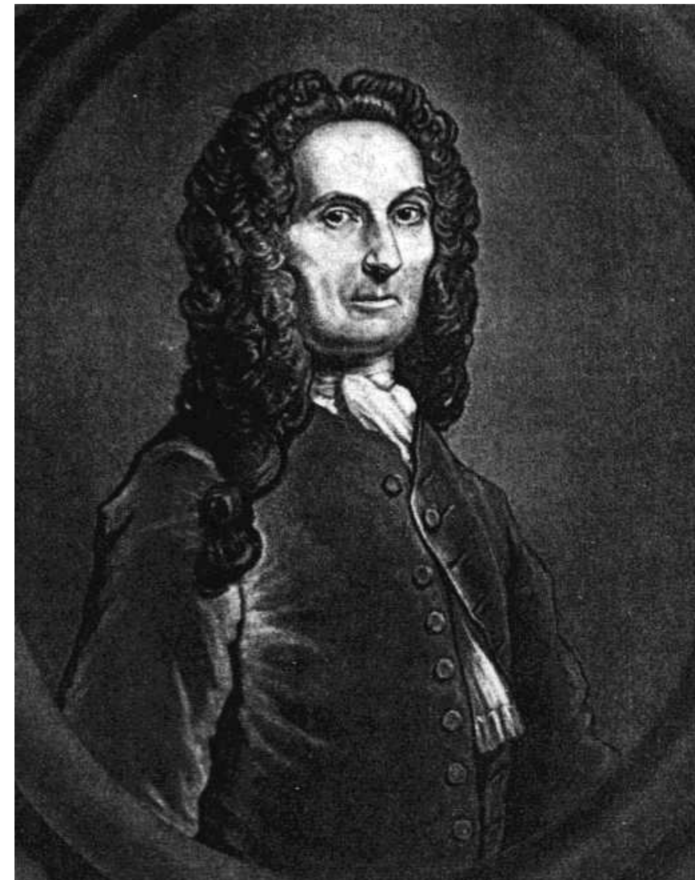
- How did people this do before computers?



MORE HISTORY

- Abraham De Moivre

  1667 - 1754



MORE HISTORY

- Abraham De Moivre

  1667 - 1754

MORE HISTORY

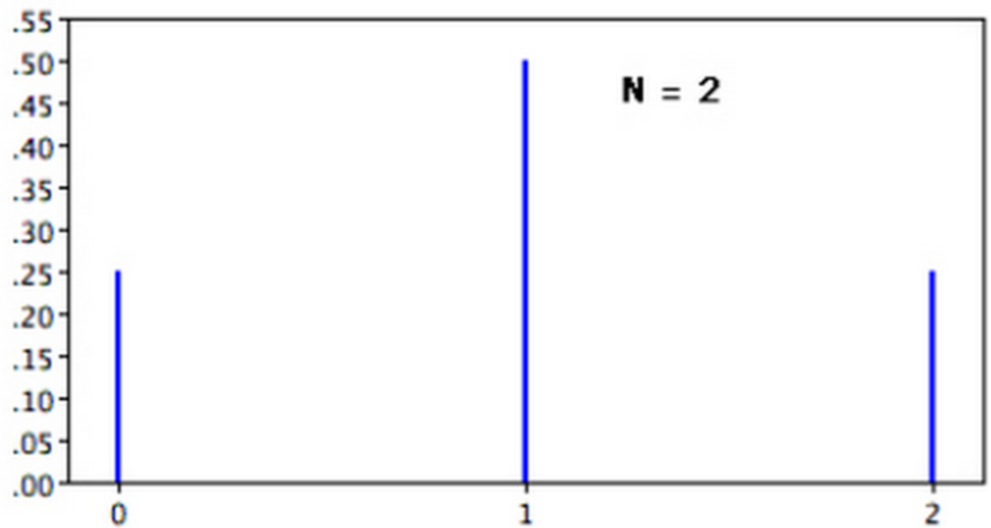
- Abraham De Moivre

  1667 - 1754

MORE HISTORY

- Abraham De Moivre
  1667 - 1754

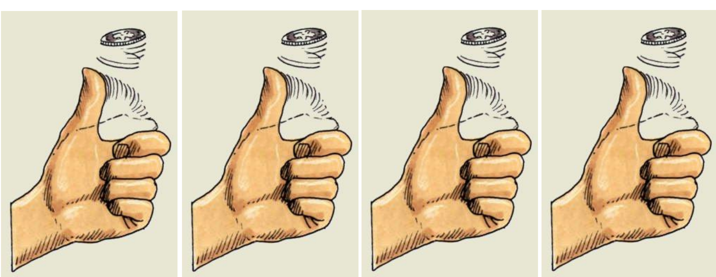


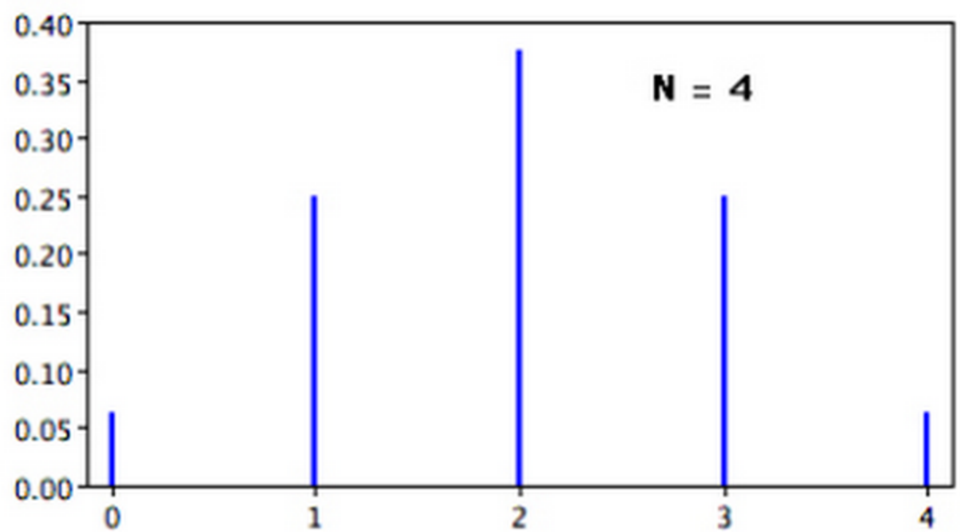


MORE HISTORY

- Abraham De Moivre

  1667 - 1754

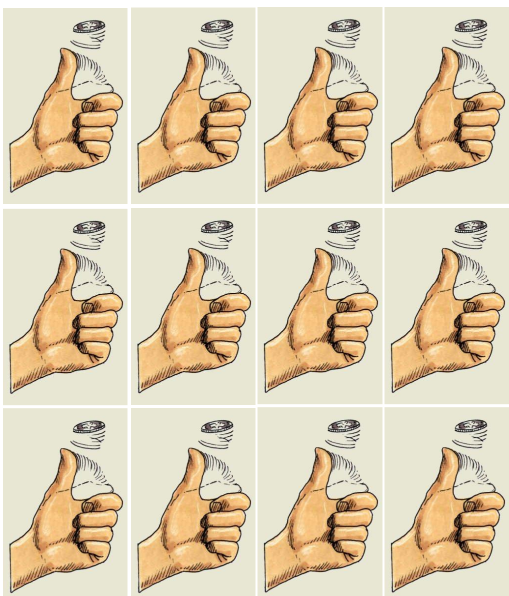


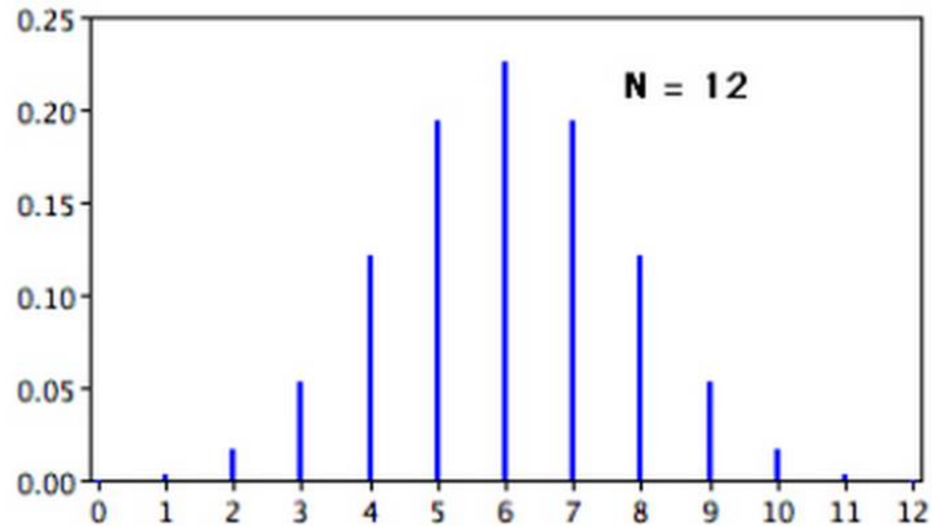


MORE HISTORY

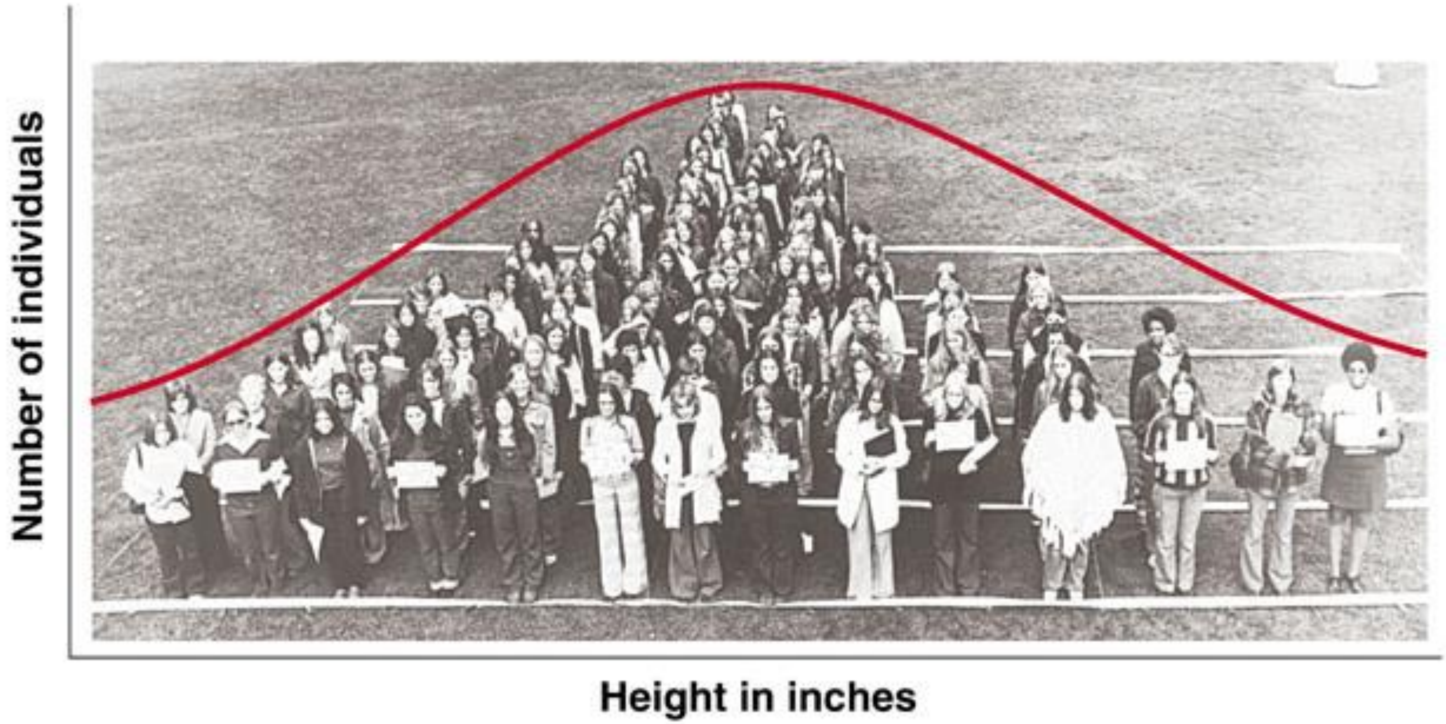
- Abraham De Moivre

  1667 - 1754

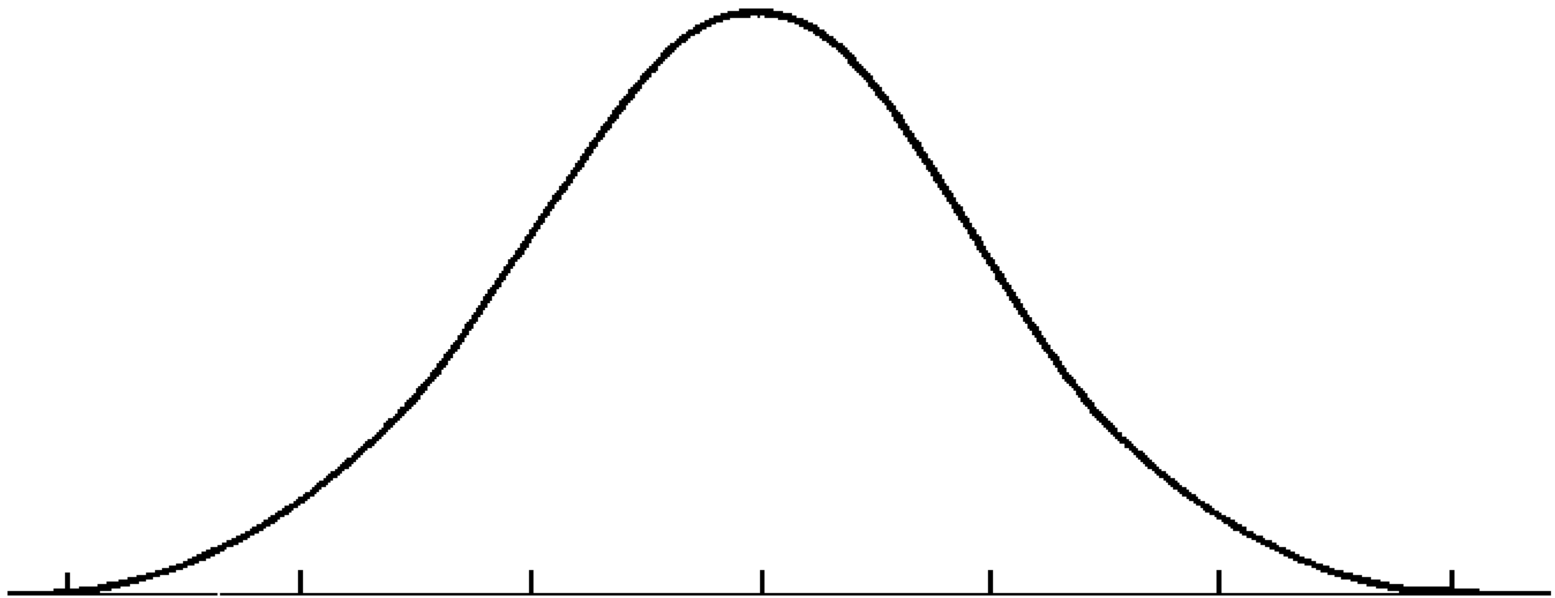




MORE HISTORY



NORMAL DISTRIBUTION



NORMAL DISTRIBUTION

- Sir Francis Galton
  1822 – 1911

Bean Machine
or Galton Board:

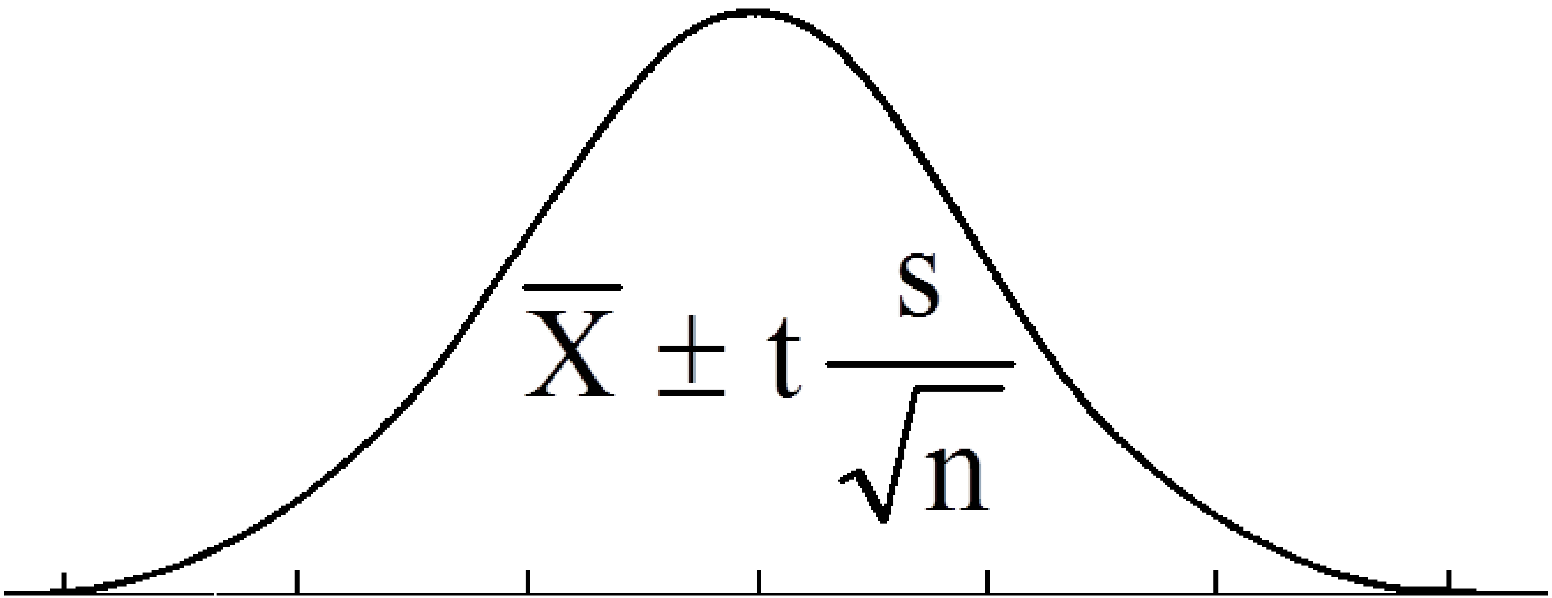
NORMAL DISTRIBUTION

Central Limit Theorem

Given certain conditions, the arithmetic mean of a sufficiently large number of iterates of independent random variables, each with a well-defined expected value and well-defined variance, will be approximately normally distributed

NORMAL DISTRIBUTION

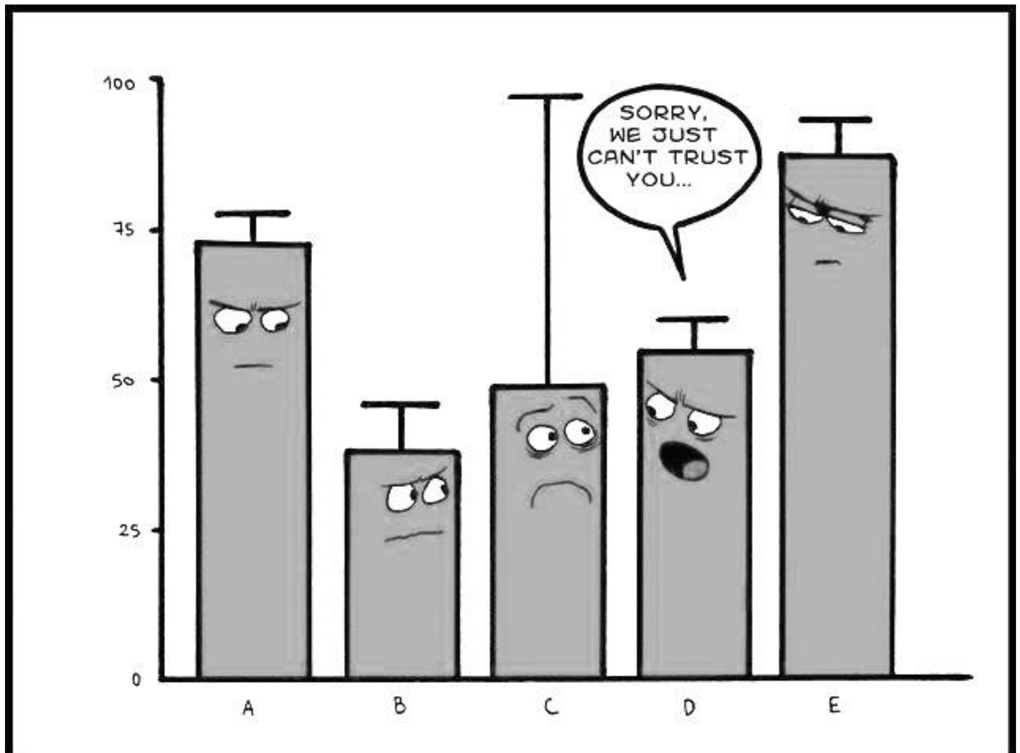
“Exact” Confidence Intervals



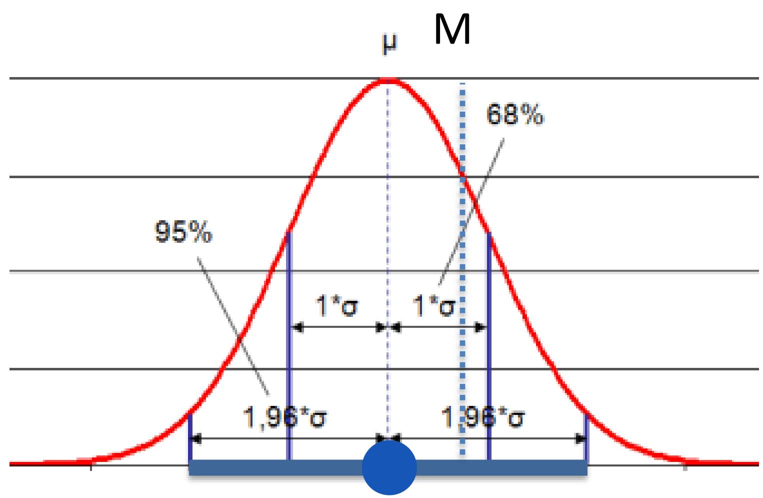
$$\overline{X} \pm t \frac{s}{\sqrt{n}}$$

t ~ 1.96 for large samples

CONFIDENCE INTERVALS

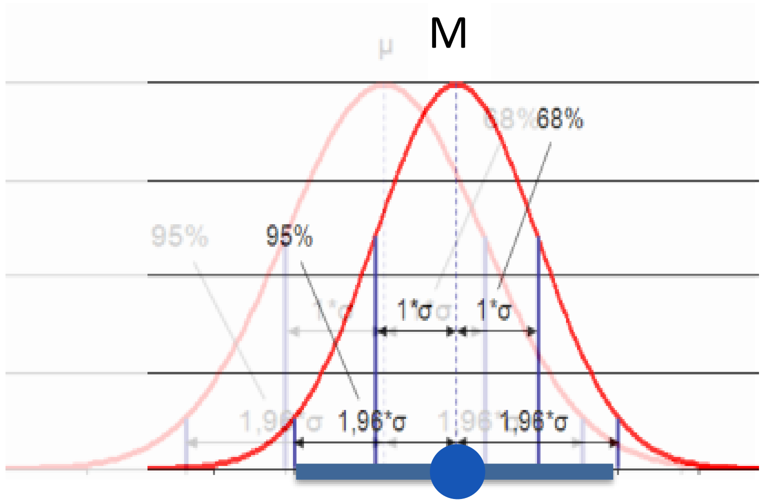


CONFIDENCE INTERVALS

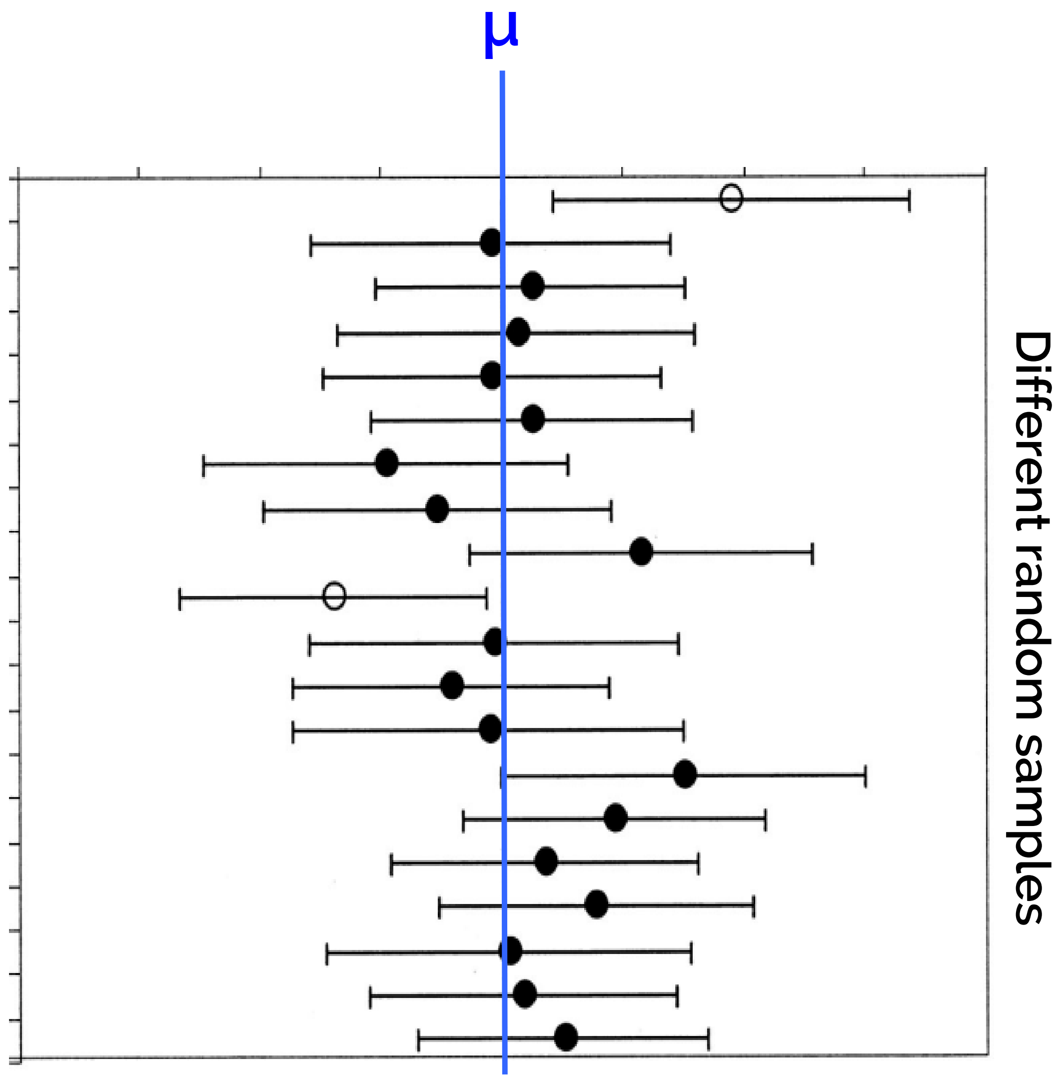


margin of error = length of blue line

CONFIDENCE INTERVALS



95% confidence interval



tinyurl.com/danceptrial2

CONFIDENCE INTERVALS

- Several interpretations
- *« a range of plausible values for μ. Values outside the CI are relatively implausible. »* (Cumming and Finch, 2005)
- Examples of presentation formats:

  2.2m, 95% CI [1.6m, 2.8m]

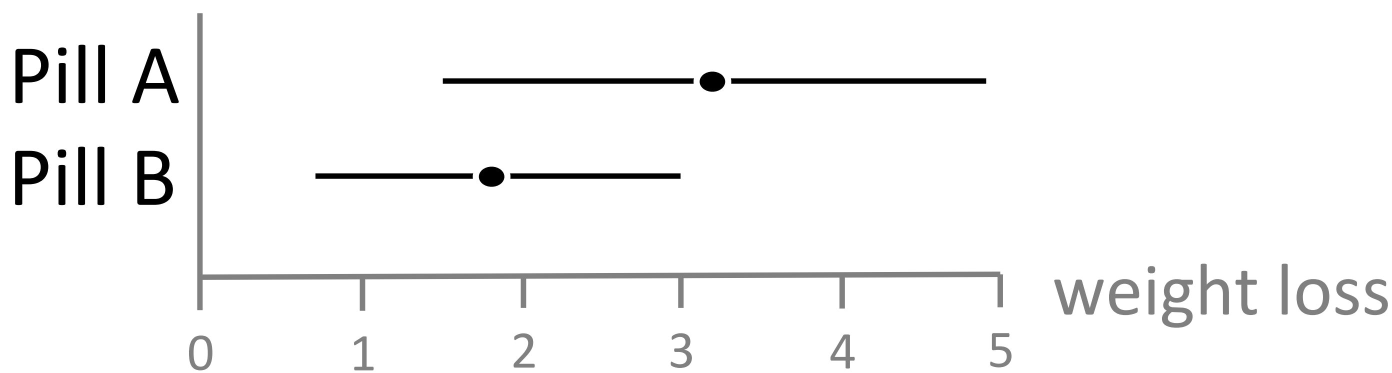
  2.2m +/- 0.6m

  from 1.6m to 2.8m

CONFIDENCE INTERVALS

« a range of plausible values for μ. Values outside the CI are relatively implausible. »

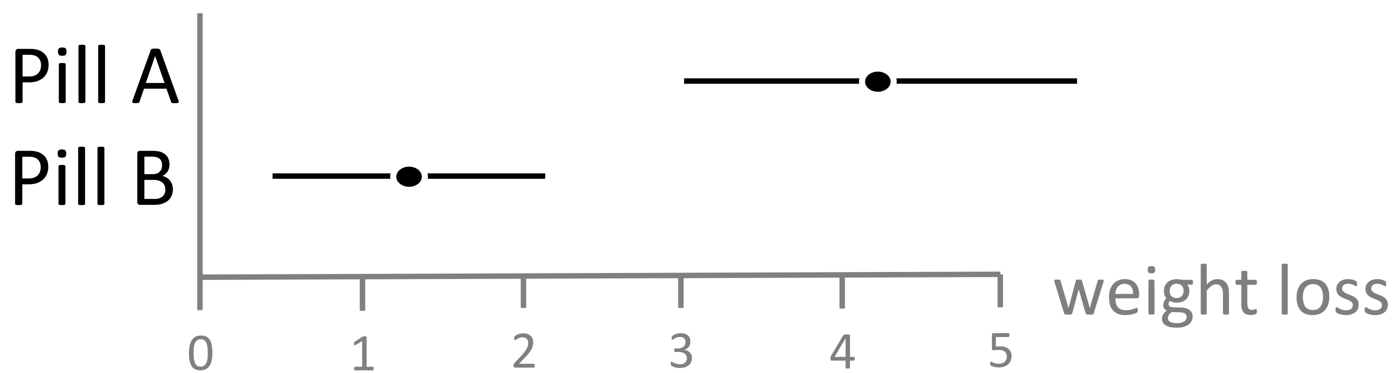
(Cumming and Finch, 2005)



CONFIDENCE INTERVALS

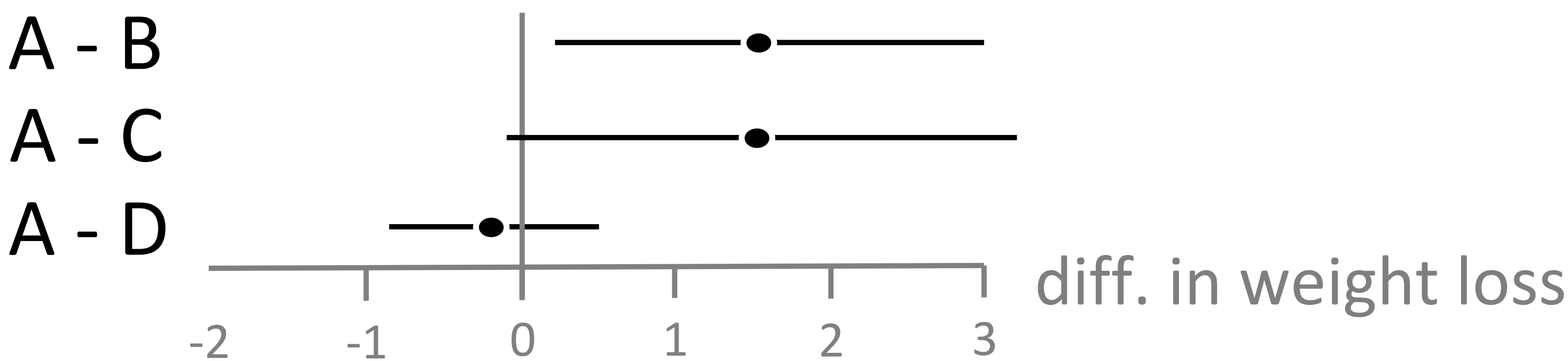
« a range of plausible values for μ. Values outside the CI are relatively implausible. »

(Cumming and Finch, 2005)



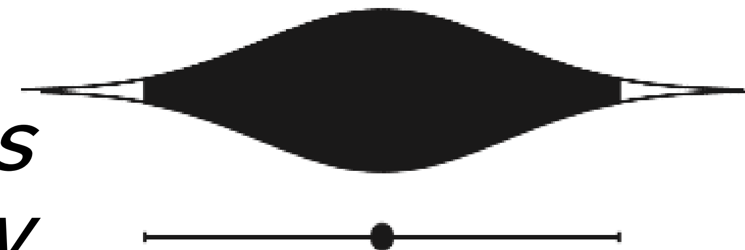
CONFIDENCE INTERVALS

- *« a range of plausible values for μ. Values outside the CI are relatively implausible. »*
  (Cumming and Finch, 2005)



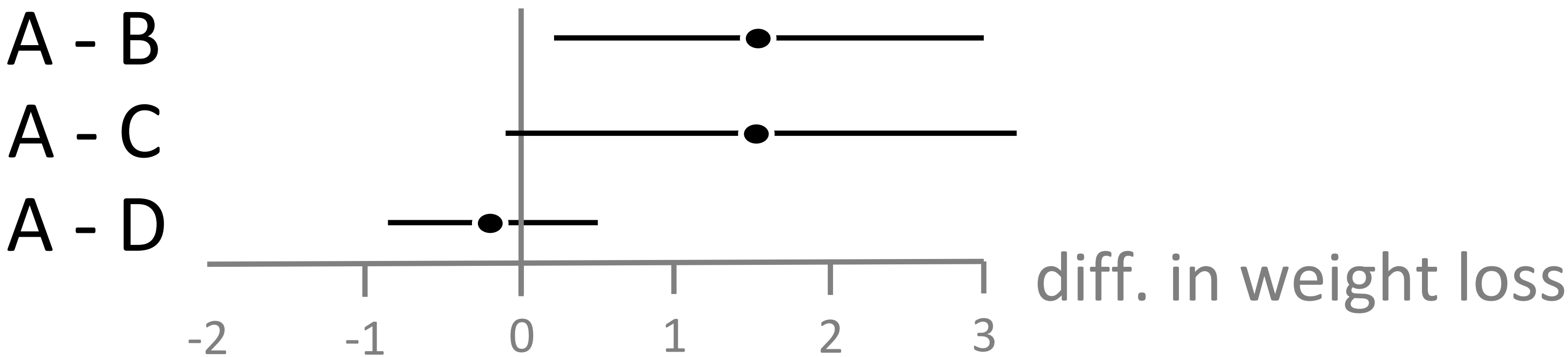
CONFIDENCE INTERVALS

"*values close to our M are the best bet for μ, and values closer to the limits of our CI are successively less good bets.*"



(Cumming, 2013)

BE VAGUE IN YOUR DESCRIPTION



"the figure provides good evidence that B outperforms A, whereas C and A seem very similar, and results are largely inconclusive concerning the difference between D and A."

PUBLISHED EXAMPLE

CAST: Effective and Efficient User Interaction for Context-Aware Selection in 3D Particle Clouds

Lingyun Yu, Konstantinos Efstathiou, Petra Isenberg, and Tobias Isenberg, *Senior Member, IEEE*

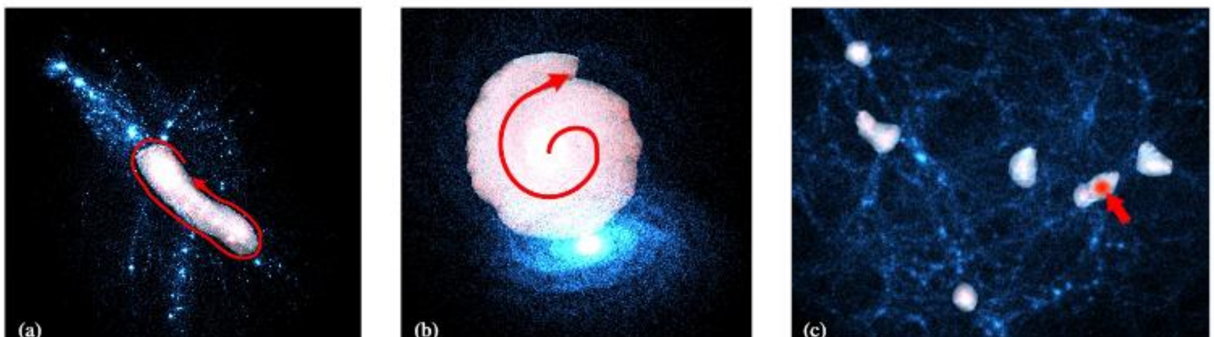


Fig. 1. (a) SpaceCast selects particle clusters by enclosing them with a lasso, based on the lasso shape; (b) TraceCast does not require an accurate lasso; and (c) with PointCast users can select tiny clusters from a noisy environment with only a single click or touch.

Abstract—We present a family of three interactive Context-Aware Selection Techniques (CAST) for the analysis of large 3D particle datasets. For these datasets, spatial selection is an essential prerequisite to many other analysis tasks. Traditionally, such interactive target selection has been particularly challenging when the data subsets of interest were implicitly defined in the form of complicated structures of thousands of particles. Our new techniques SpaceCast, TraceCast, and PointCast improve usability and speed of spatial selection in point clouds through novel context-aware algorithms. They are able to infer a user's subtle selection intention from gestural input, can deal with complex situations such as partially occluded point clusters or multiple cluster layers, and can all be fine-tuned after the selection interaction has been completed. Together, they provide an effective and efficient tool set for the fast exploratory analysis of large datasets. In addition to presenting Cast, we report on a formal user study that compares our new techniques not only

PUBLISHED EXAMPLE

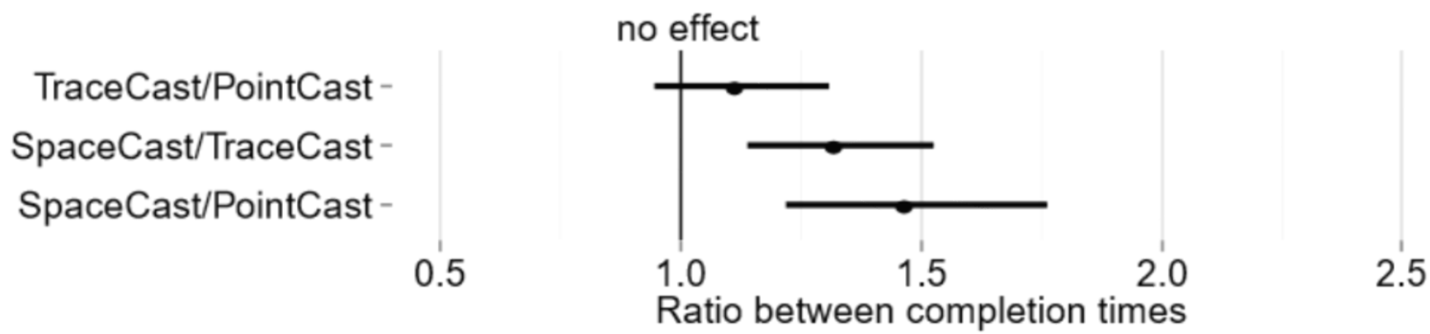
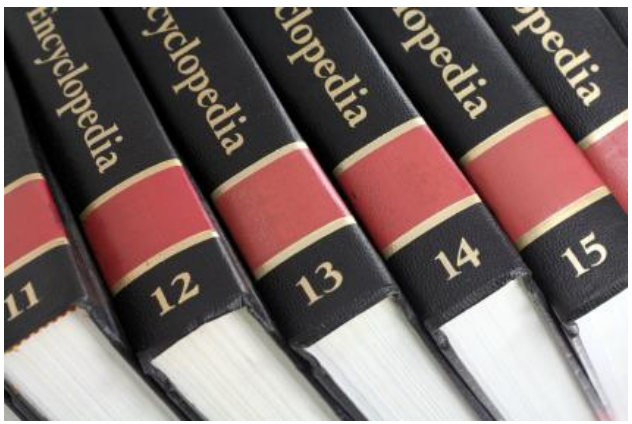


Fig. 6. Ratios between mean completion times for the Cast selection techniques. Error bars show 95% confidence intervals.

Thus, overall we have good evidence that both PointCast and TraceCast outperform SpaceCast, and some indication that PointCast may outperform TraceCast. At any rate, the differences among Cast methods are marginal compared to the differences between each Cast method and CloudLasso or CylinderSelection.

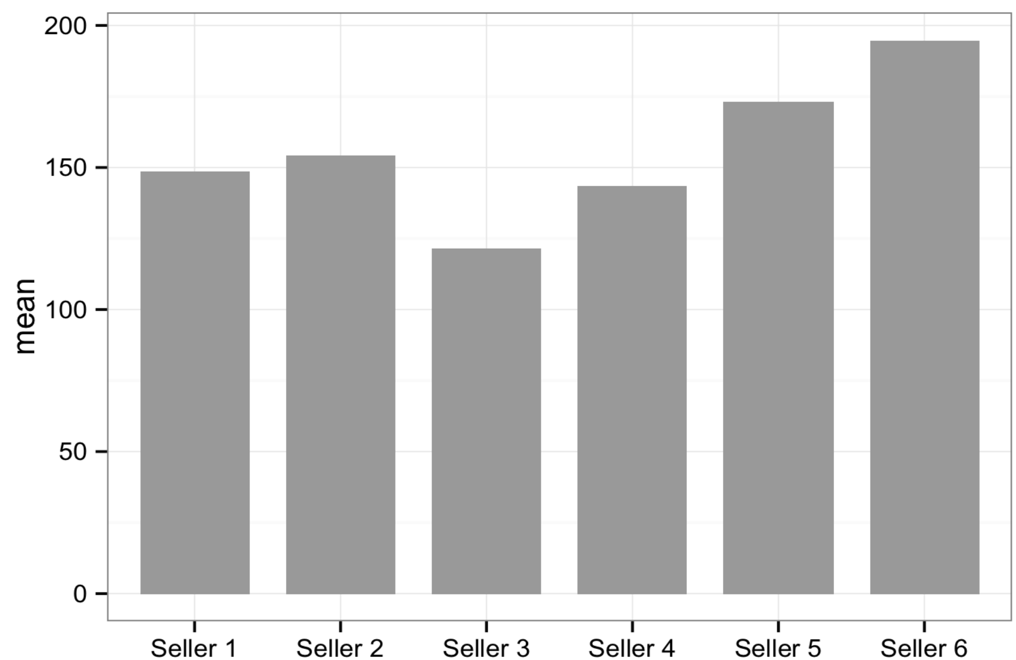
BACK TO OUR EXAMPLE

- Selling encyclopedias



Average Sales

Seller 1	Seller 2	Seller 3	Seller 4	Seller 5	Seller 6
€149	€154	€122	€143	€173	€195



Fair Statistical Communication in HCI

Pierre Dragicevic

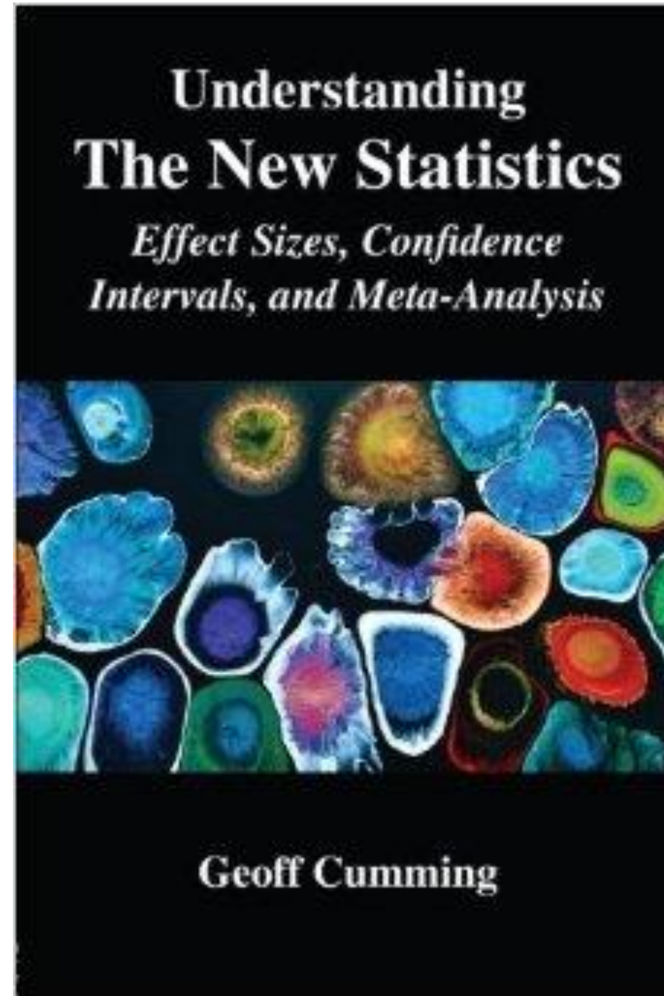
Preprint v.1.6.3, to appear in February 2016. How to cite:


```
@incollection{dragicevic2016fair,
  author = {Pierre Dragicevic},
  title = {Fair Statistical Communication in {HCI}},
  editor = {J. Robertson and M.C. Kaptein},
  booktitle = {Modern Statistical Methods for {HCI}},
  publisher = {Springer},
  year = {2016},
  note = {In press}}
```


Abstract Statistics are tools to help end users accomplish their task. In research, to be qualified as usable, statistical tools should help researchers advance scientific knowledge by supporting and promoting the effective communication of research findings. Yet areas such as human-computer interaction (HCI) have adopted tools — i.e., *p*-values and dichotomous testing procedures — that have proven to be poor at supporting these tasks. The abusive use of these procedures has been severely criticized in a range of disciplines for several decades, suggesting that tools should be blamed, not end users. This chapter explains in a non-technical manner why it would be beneficial for HCI to switch to an *estimation* approach, i.e., reporting informative charts with effect sizes and interval estimates, and offering nuanced interpretations of our results. Advice is offered on how to communicate our empirical results in a clear, accurate, and transparent way without using any tests or *p*-values.

1 Introduction

A common analogy for statistics is the toolbox. As it turns out, researchers in human-computer interaction (HCI) study computer tools. A fairly uncontroversial position among them is that tools should be targeted at end users, and that we should



Bad Stats: Not What It Seems

Towards a Statistical Reform in HCI and Visualization

Pierre Dragicevic and colleagues

This page provides arguments and reading material to explain why it would be beneficial for human-computer interaction (HCI) and information visualization (infovis) to stop doing mindless null hypothesis significance testing (NHST) and start reporting informative charts with effect sizes and interval estimates, as well as offering more nuanced interpretations of our results. Our scientific standards can also be greatly improved by planning analyses and sharing experimental material online.

Content:

Fair Statistical Communication in HCI (book chapter)
Bad Stats are Miscommunicated Stats (BELIV 2014 Keynote)
Running an HCI Experiment in Multiple Parallel Universes (Alt.CHI 2014 Paper)
Quotes about null hypothesis significance testing (NHST)
Links
Reading List
More Readings
Papers (somehow) in favor of NHST
Papers against confidence intervals
Papers from the HCI Community

http://tinyurl.com/stats-dresden