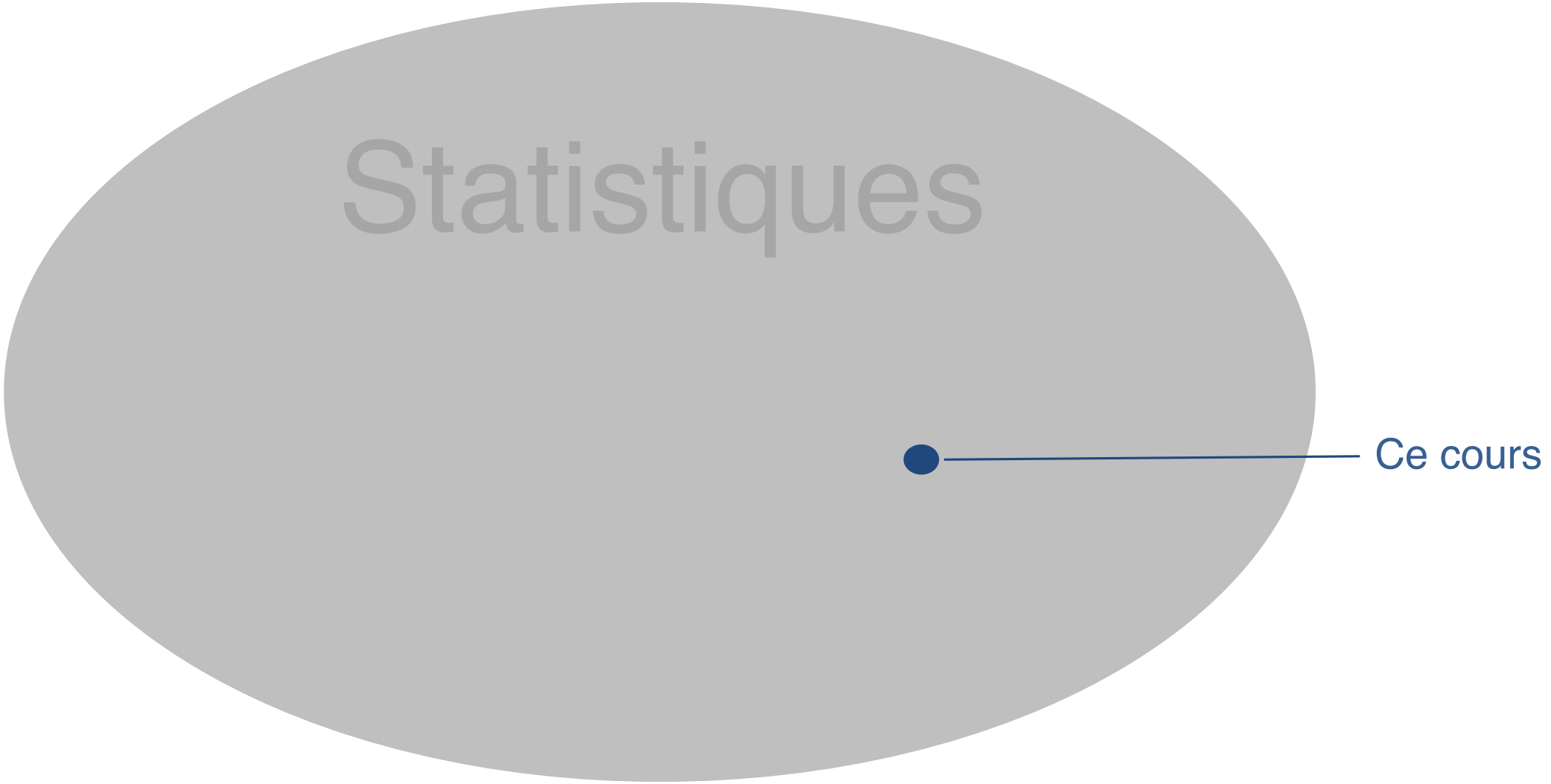
INTRODUCTION AUX NOUVELLES STATISTIQUES POUR L'IHM

Pierre Dragicevic

7èmes Rencontres des Jeunes Chercheurs en IHM
Juin 2015

CONTENU DE CE COURS



OBJECTIFS

- Acquérir les intuitions et la terminologie de base sur les stats
- Première exposition à R
- Accent sur les aspects haut-niveau
- Sensibilisation aux abus des statistiques
- Accent sur les "nouvelles statistiques"

ORGANISATION

- Partie I - notions élémentaires de stats
- Partie II - analyses préliminaires en R
- Partie III - bien utiliser les stats en IHM

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.

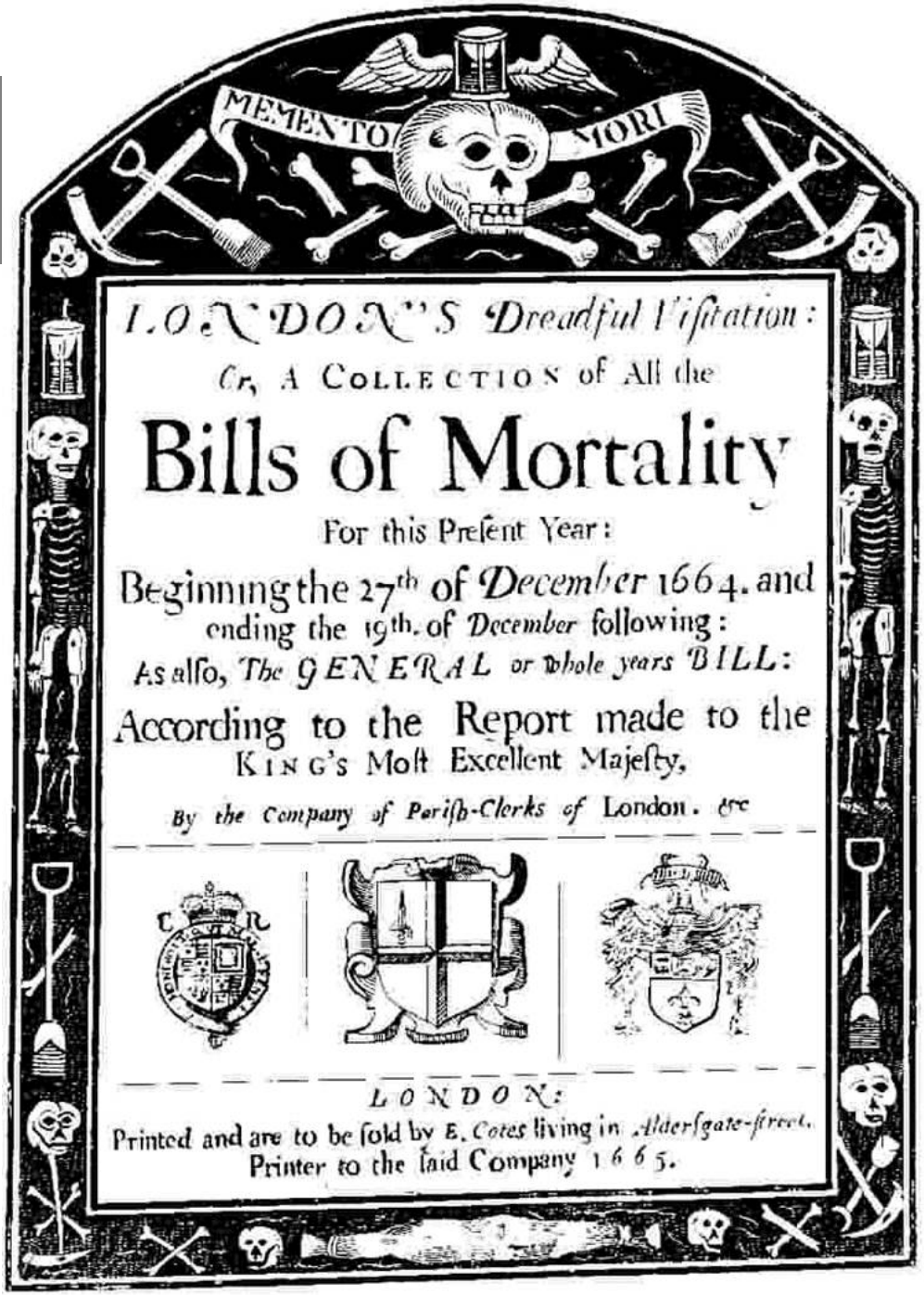
ORIGINS

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

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	165	1656	1657	1658	1659	1660	1629	1630	1631	1632	163	1634	1635	1636	1629 1630 1631 1632	1633 1634 1635 1636	1647 1648 1649 1650	1651 1652 1653 1654	1655 1656 1657 1658	1629 1640 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	50	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	70	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	953	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	2	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	16	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	81	101	85	120	113	179	116	167	48	57					37	50	105	87	341	359	497	247	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	2216	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	211	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	559	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	49	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	819	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	2[illegible]	34	27	29	243
French-Pox	18	29	15	18	21	20	20	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	15	9	14	16	24	18	11	36	8	8	6	15		3	8	7	37	18	48	47	72	32	222
Aven-Evil		1	11	2		2	0	0	5	3	4	5	35	26							4	2	0	0	14	14	17	46	0,4
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	06
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	13	18	6	11	7	11	8	13	6	7	13	5	14	14	6	11	6	5	4	2	2	5	28	13	47	39	31	20	161

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	1
Colick, and Wind	103	71	85	82	76	102	84	101	85	120	
Cold, and Cough							41	36	21	58	
Conſumption, and Cough	2423	2200	2388	1988	2350	2410	2216	2868	2606	3184	2
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	33	30	43	41	
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	1150	184	525	1279	839	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	11	9	7	7	5	6	8	7	
Grief	12	13	16	7	17	14	11	17	10	13	
Hanged and made										16	

ORIGINS

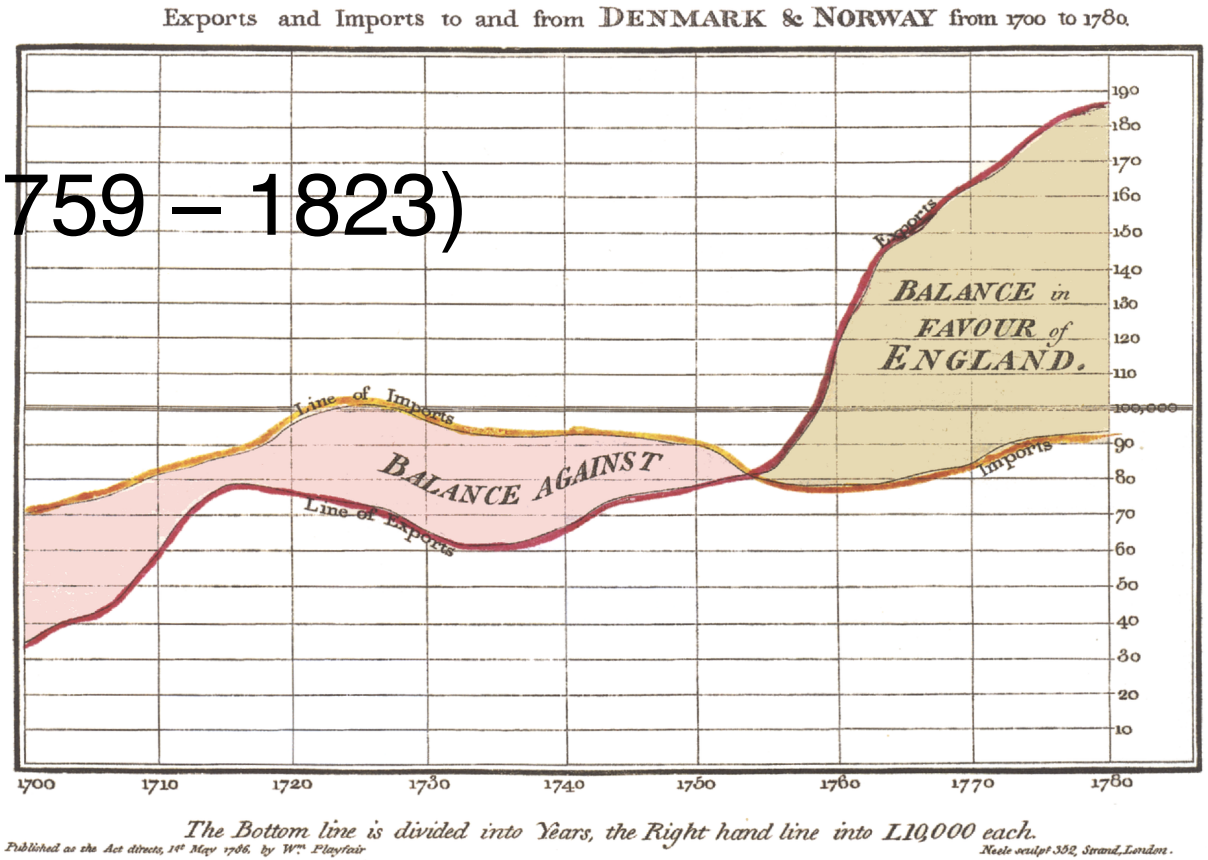
- John Graunt, 1662
  Observations on the bills of mortality
  - First “life tables”
  - Dispelled several myths about the plague
  - First analysis of sex ratio
  - First realistic estimate of the population in London

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
  - ...

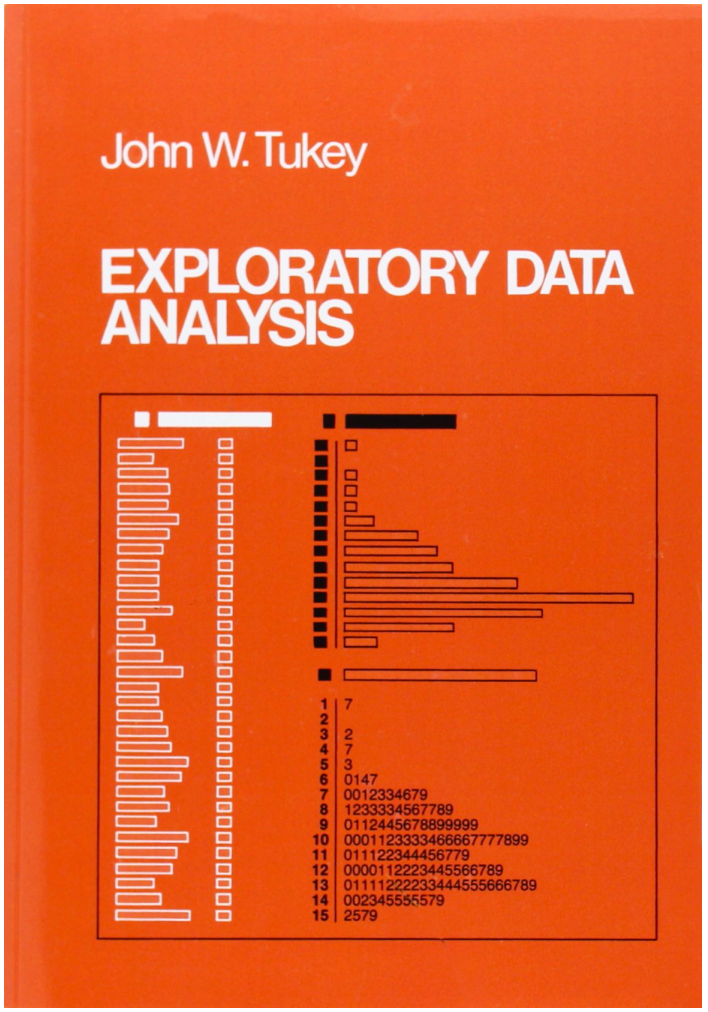
STATS & VISUALIZATION

- Statistical Charts
  - William Playfair (1759 – 1823)

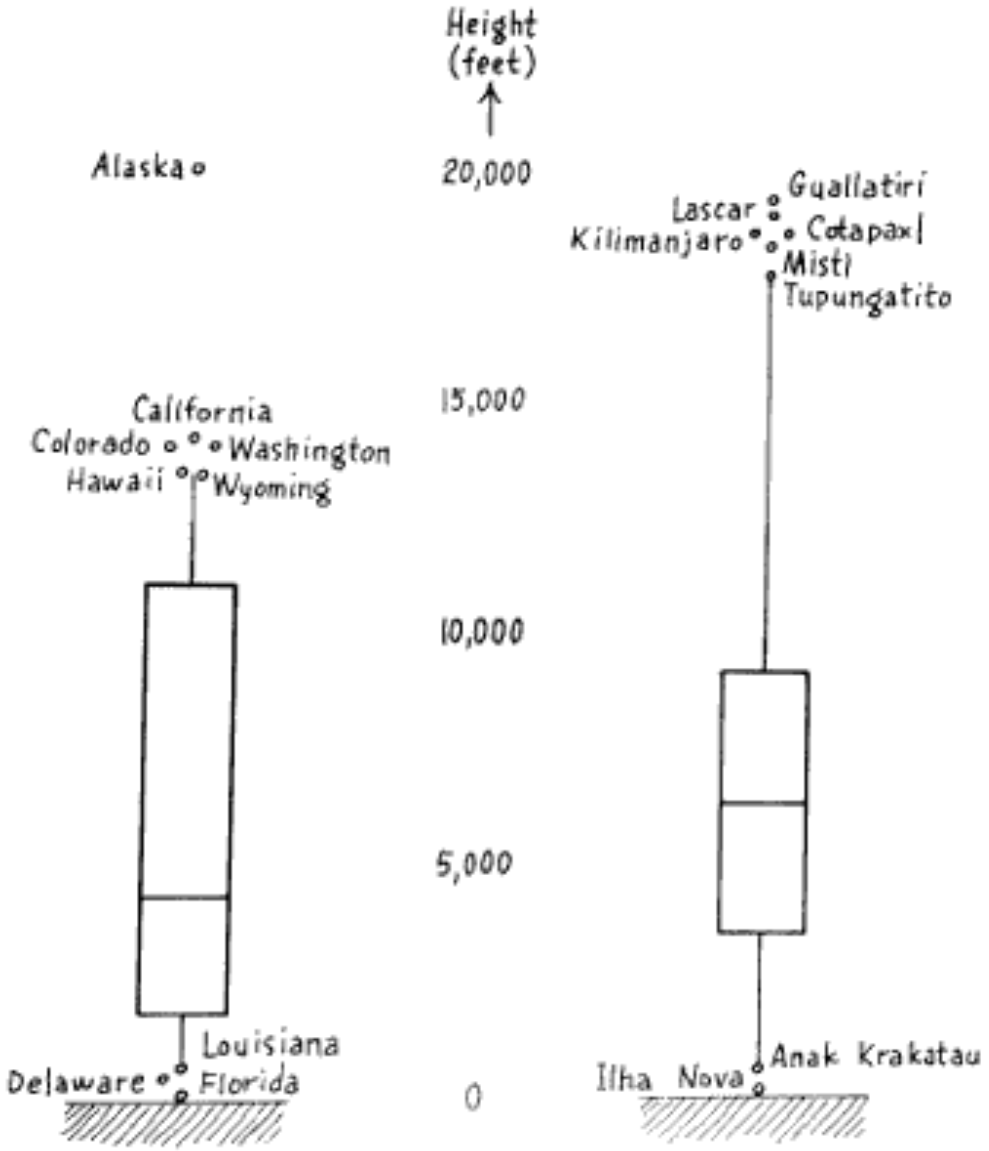


STATS & VISUALIZATION

- Exploratory Data Analysis
  - Tukey, 1977



Box-and-whisker plots with end values identified



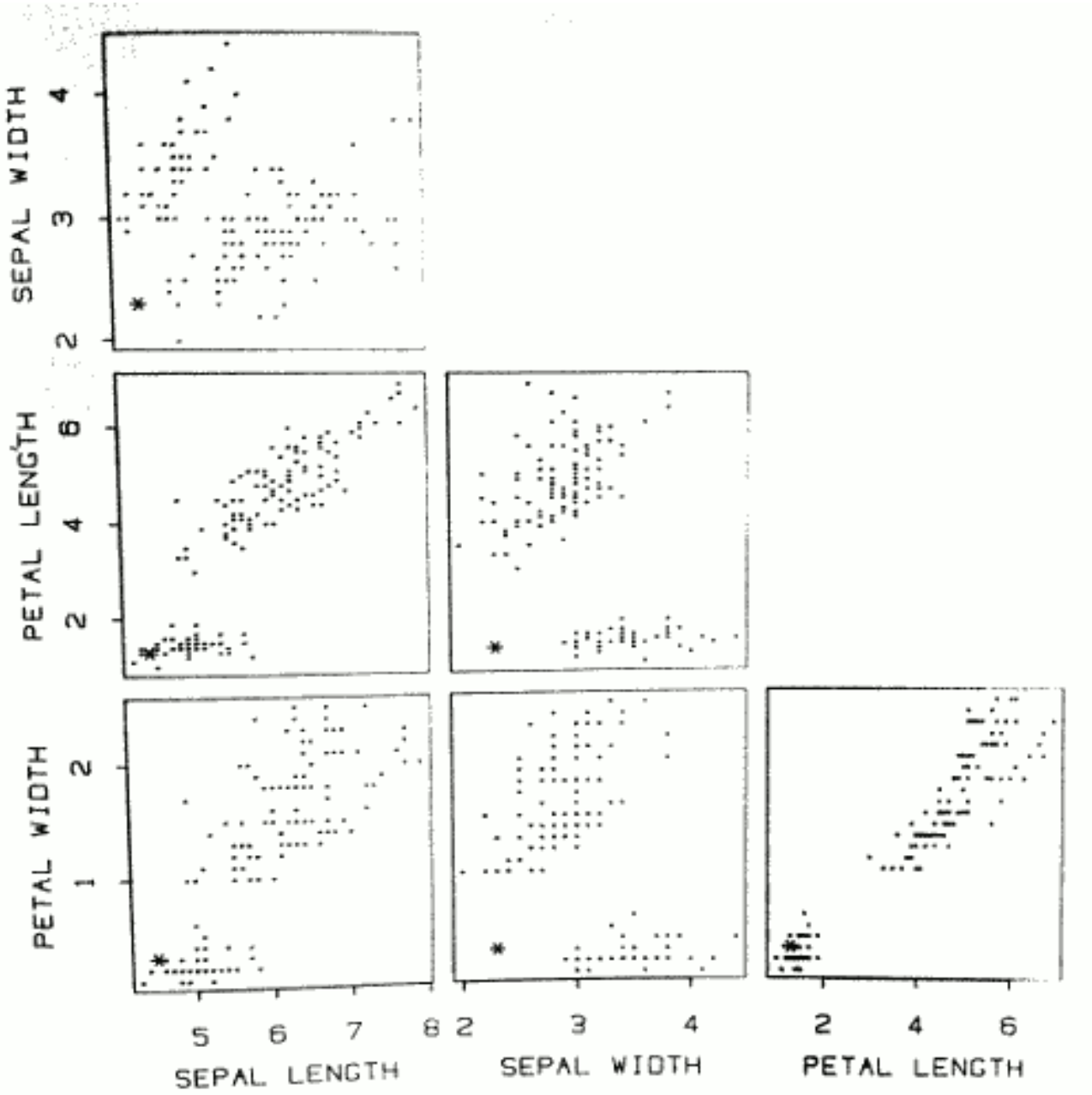
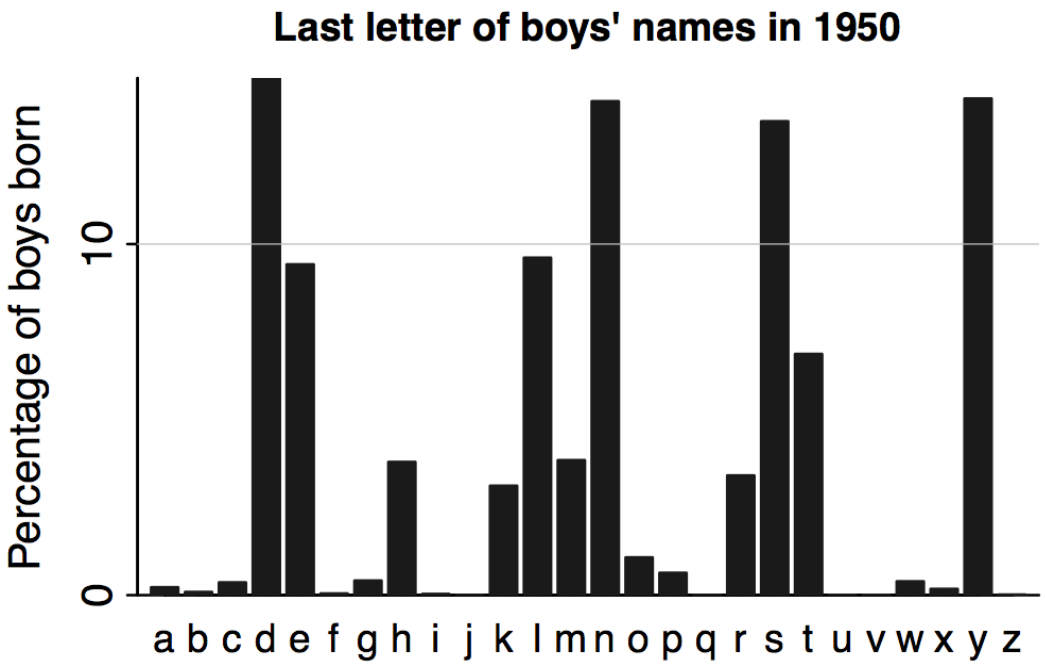
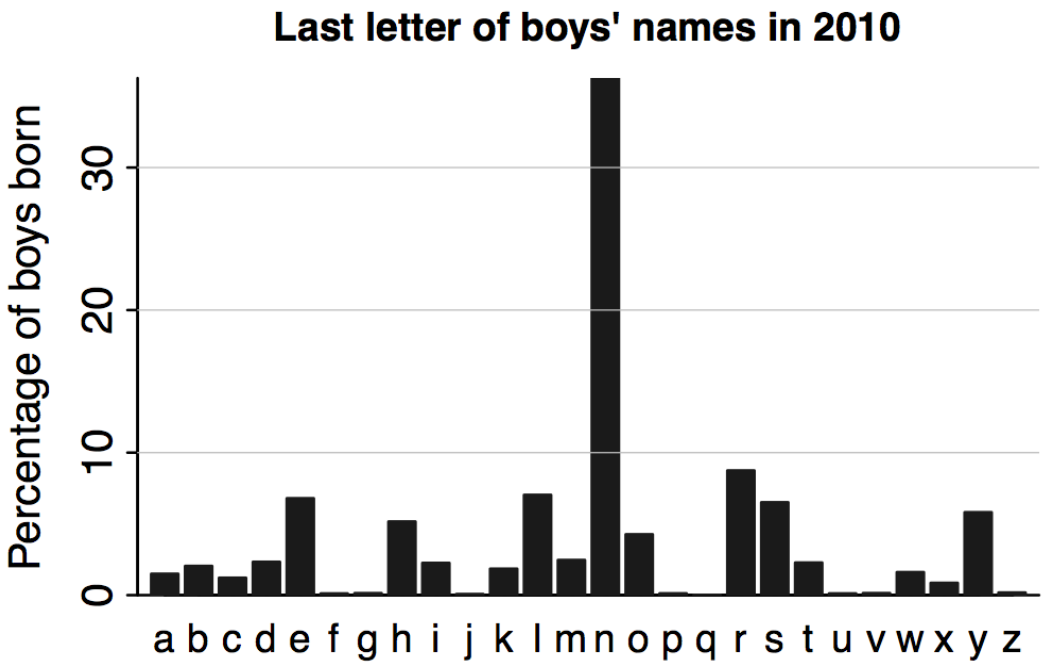
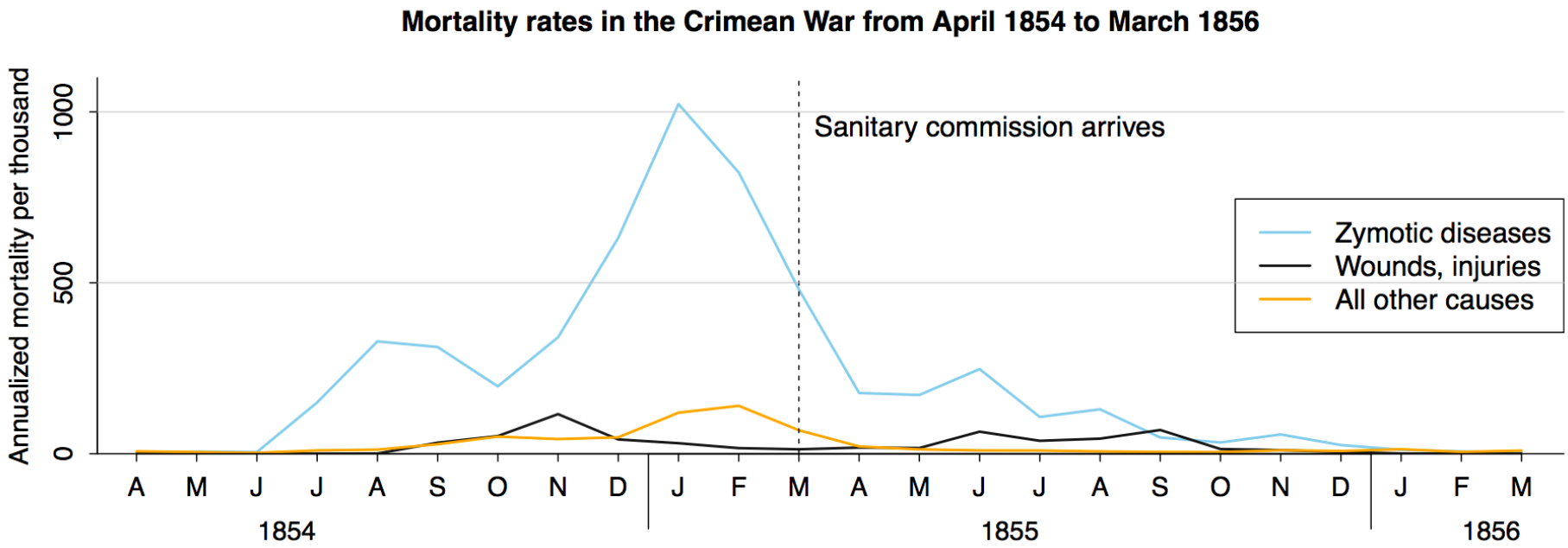
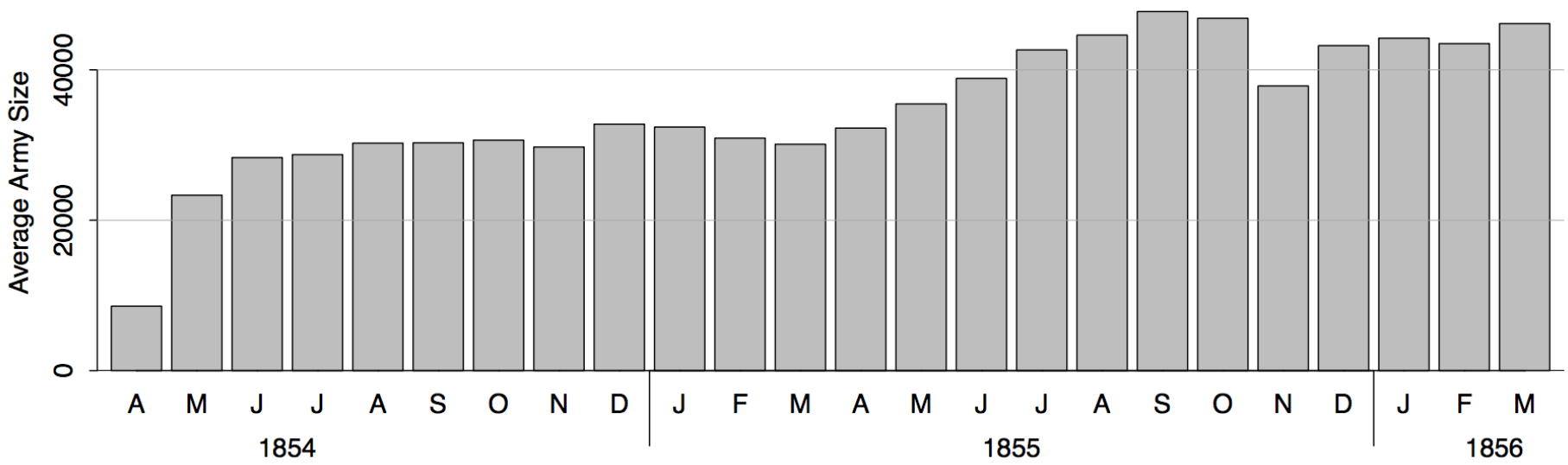


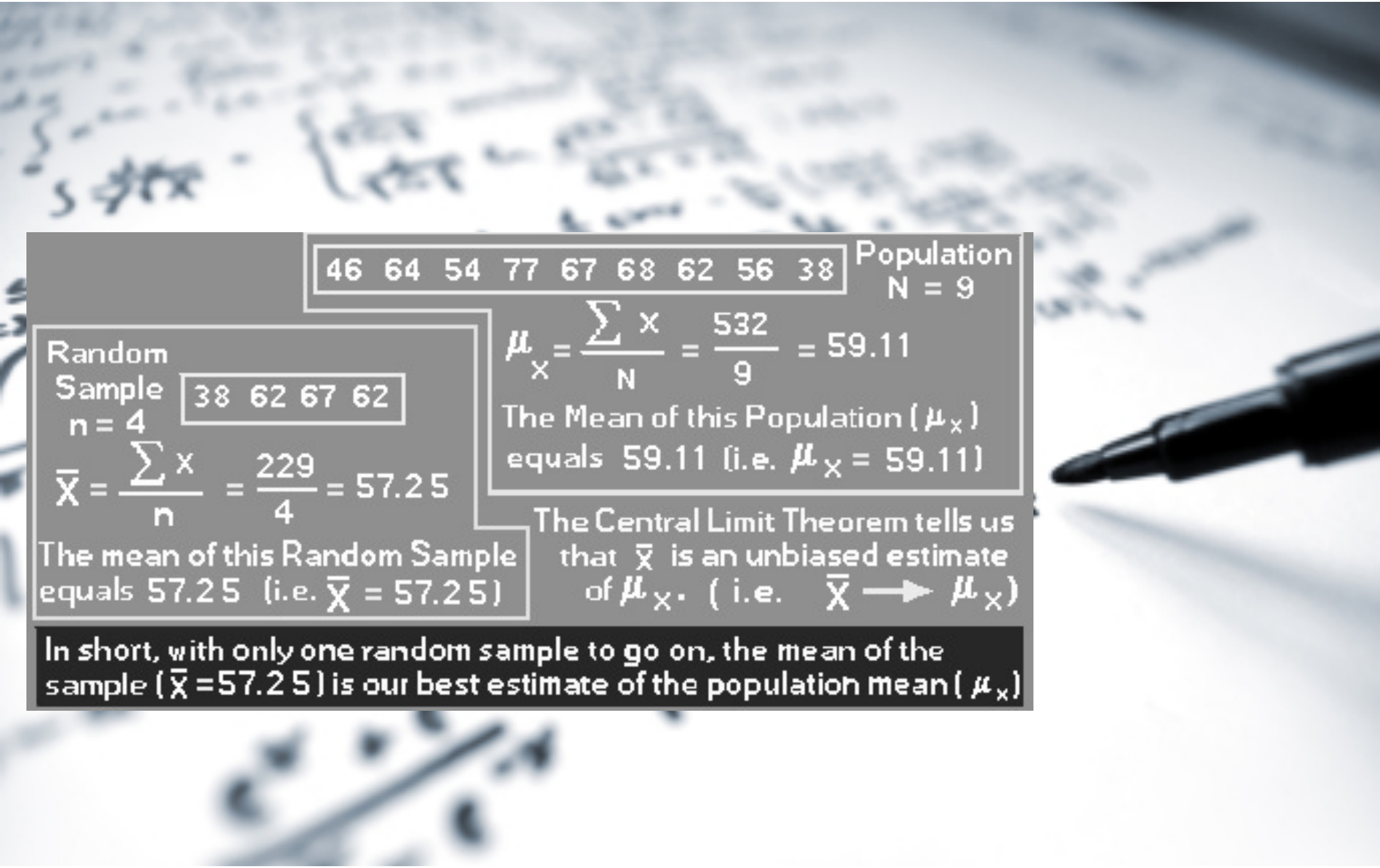
Figure 5.14 Generalized draftsman's display of the four-dimensional iris data (like Figure 5.11), with one flower plotted as an asterisk.

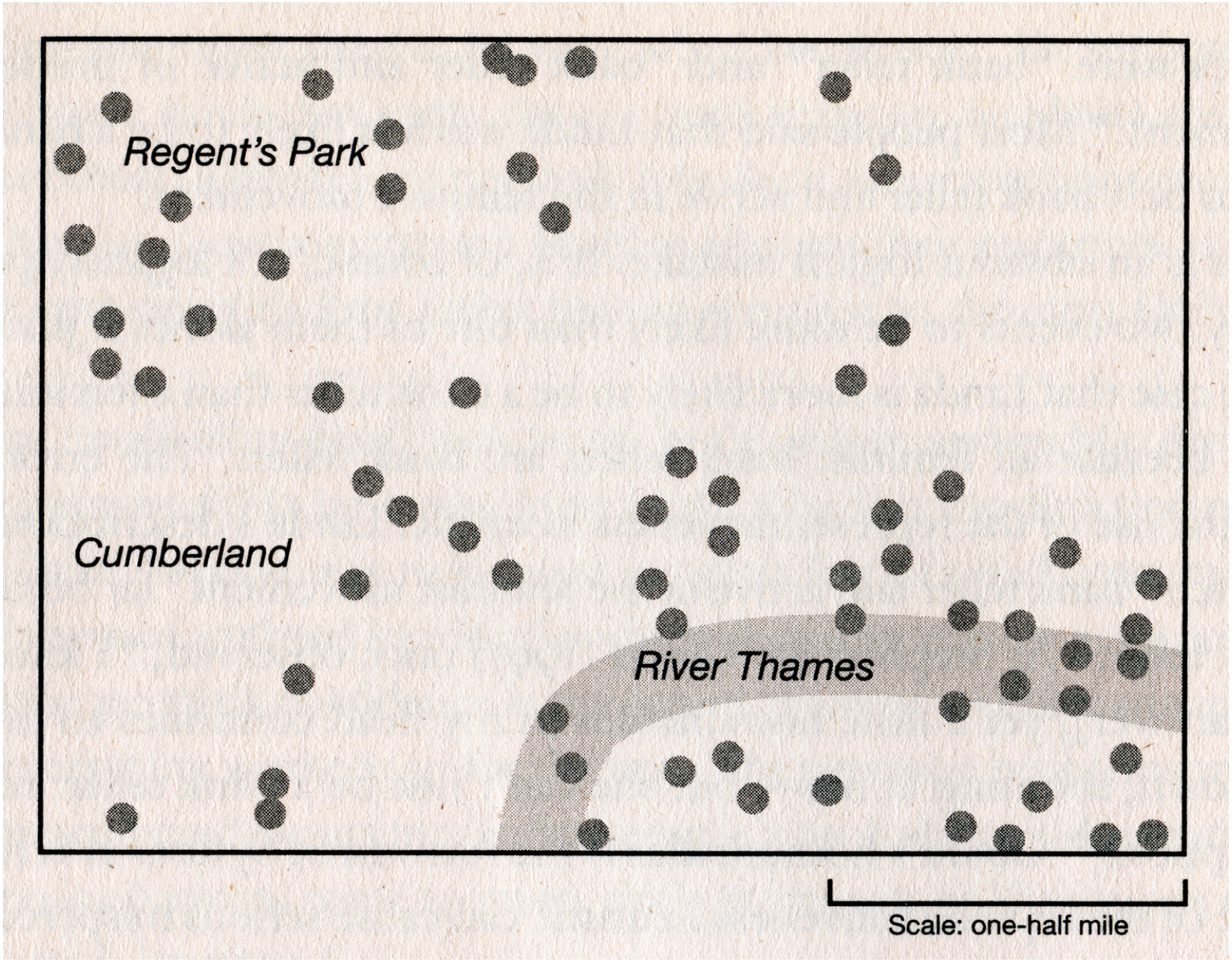

Last letter of boys' names in 1950
Percentage of boys born
0
10
a b c d e f g h i j k l m n o p q r s t u v w x y z


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

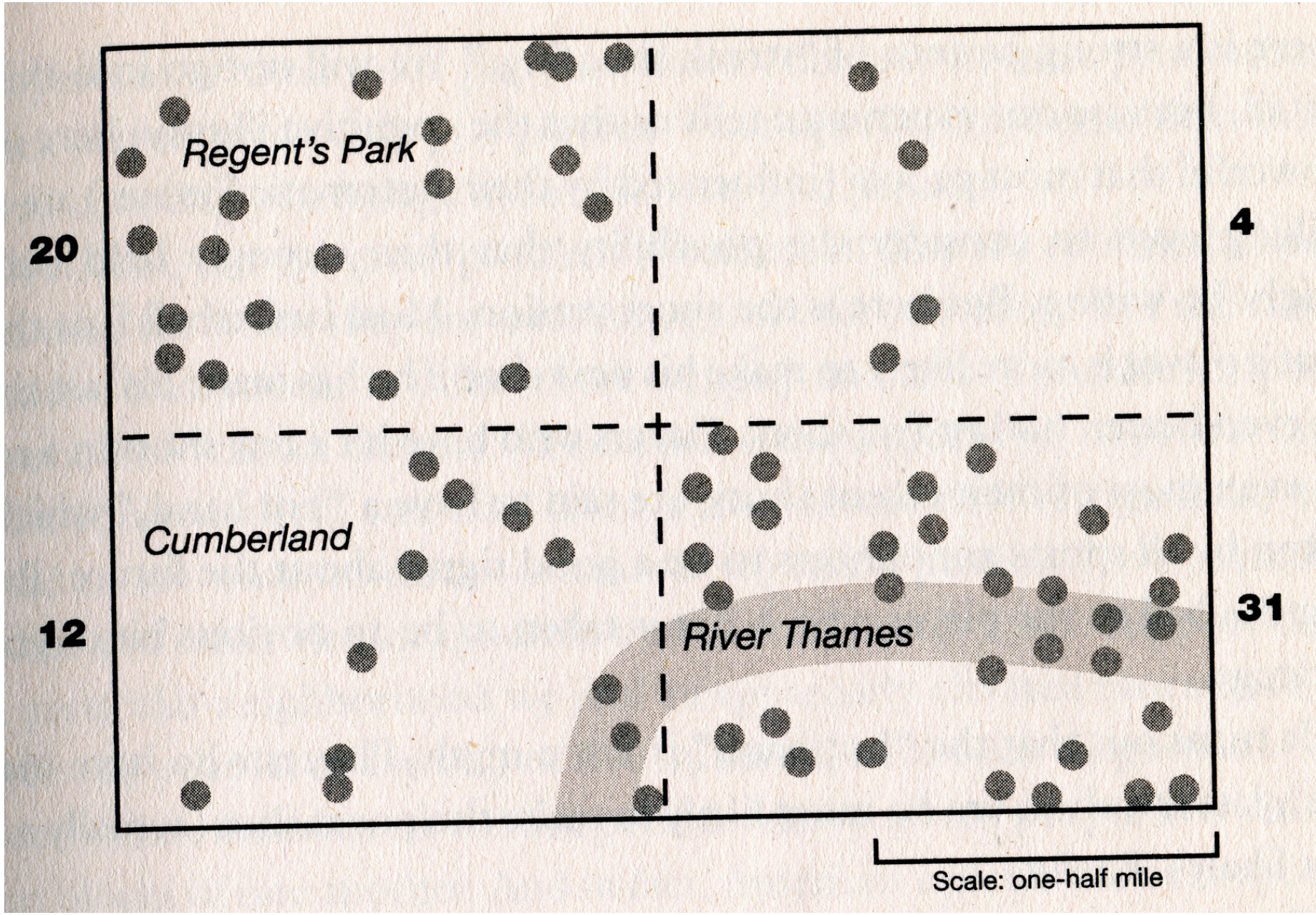

Mortality rates in the Crimean War from April 1854 to March 1856
Annualized mortality per thousand
0
500
1000
Sanitary commission arrives
Zymotic diseases
Wounds, injuries
All other causes
A M J J A S O N D J F M A M J J A S O N D J F M
1854
1855
1856


British Army Size in the Crimean War from April 1854 to March 1856
Average Army Size
0
20000
40000
A M J J A S O N D J F M A M J J A S O N D J F M
1854
1855
1856


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. $\mu_X = 59.11$)
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. $\bar{X} \longrightarrow \mu_X$)
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

STATS & VISUALIZATION

- Confirmatory Analysis
  - Testing hypotheses
  - Example: is this new drug effective?
  - Strong focus on automatic procedures, computation and objectivity
  - Looking at data can impair objectivity:
    - Data dredging, snooping, fishing, mining

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).

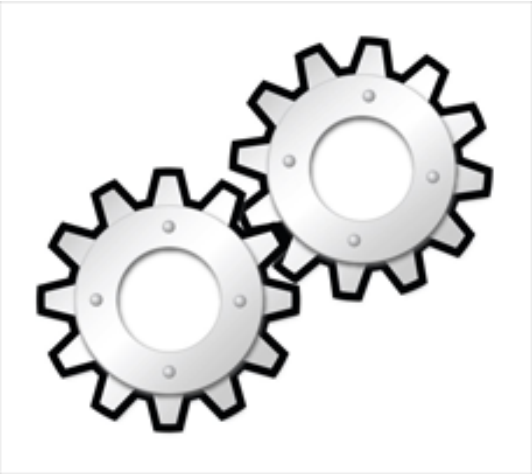
Quoted from the SAS Institute

WHAT ARE STATS?

- A set of tools and methods
- Old tradition:
  - Origins in demographics
  - Draws from mathematics & probability theory
  - Visual representations are also important
  - A (generally) strong focus on (computationally cheap) numerical calculations

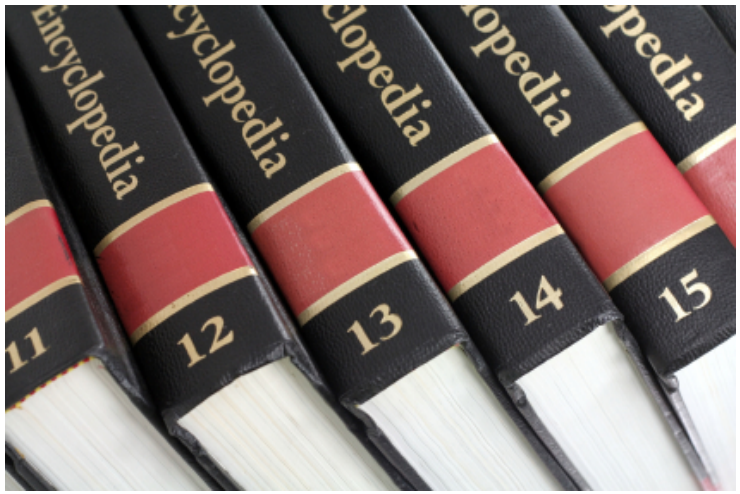
STATISTICAL TOOLS

DESCRIPTIVE STATISTICS



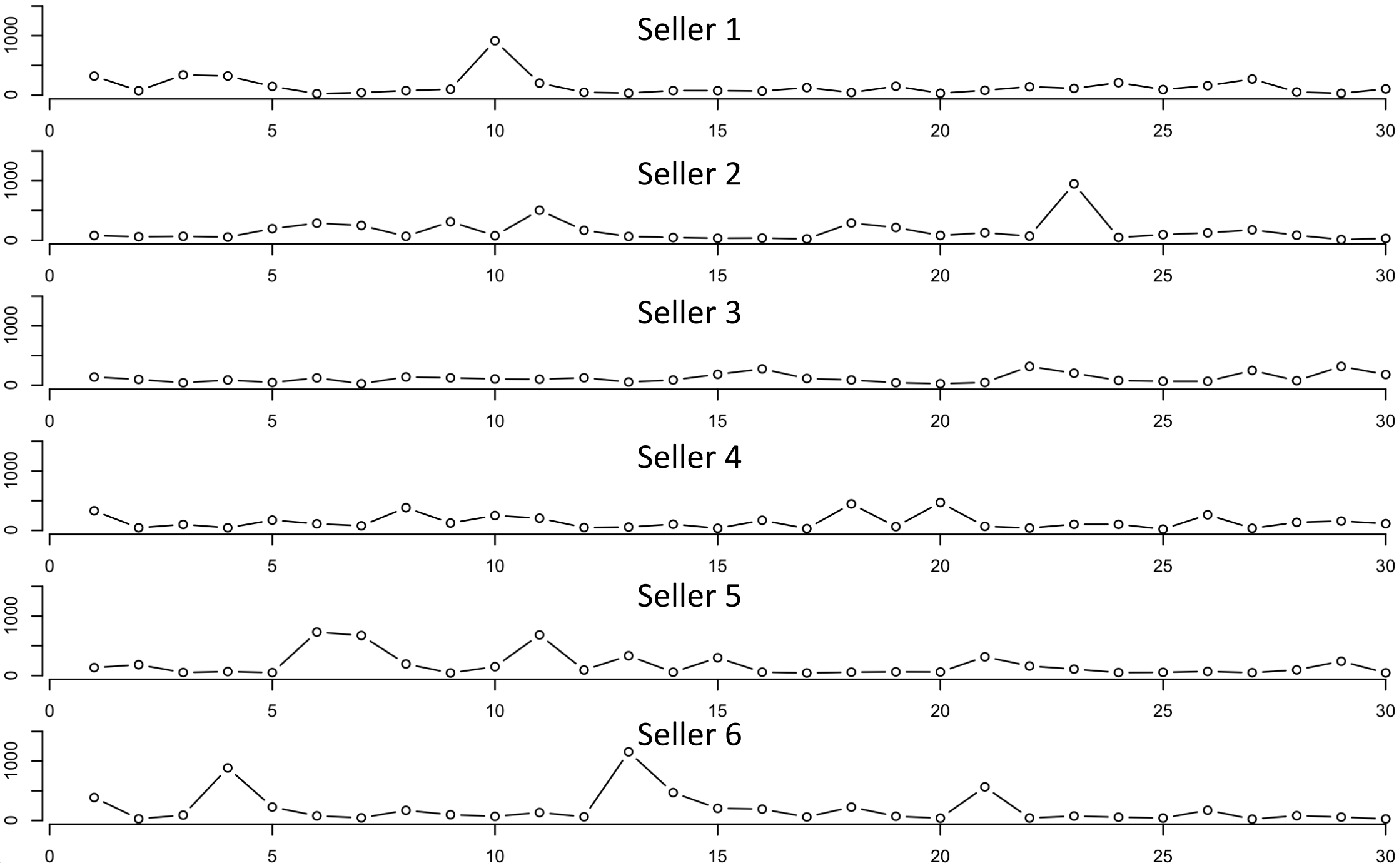
AN EXAMPLE

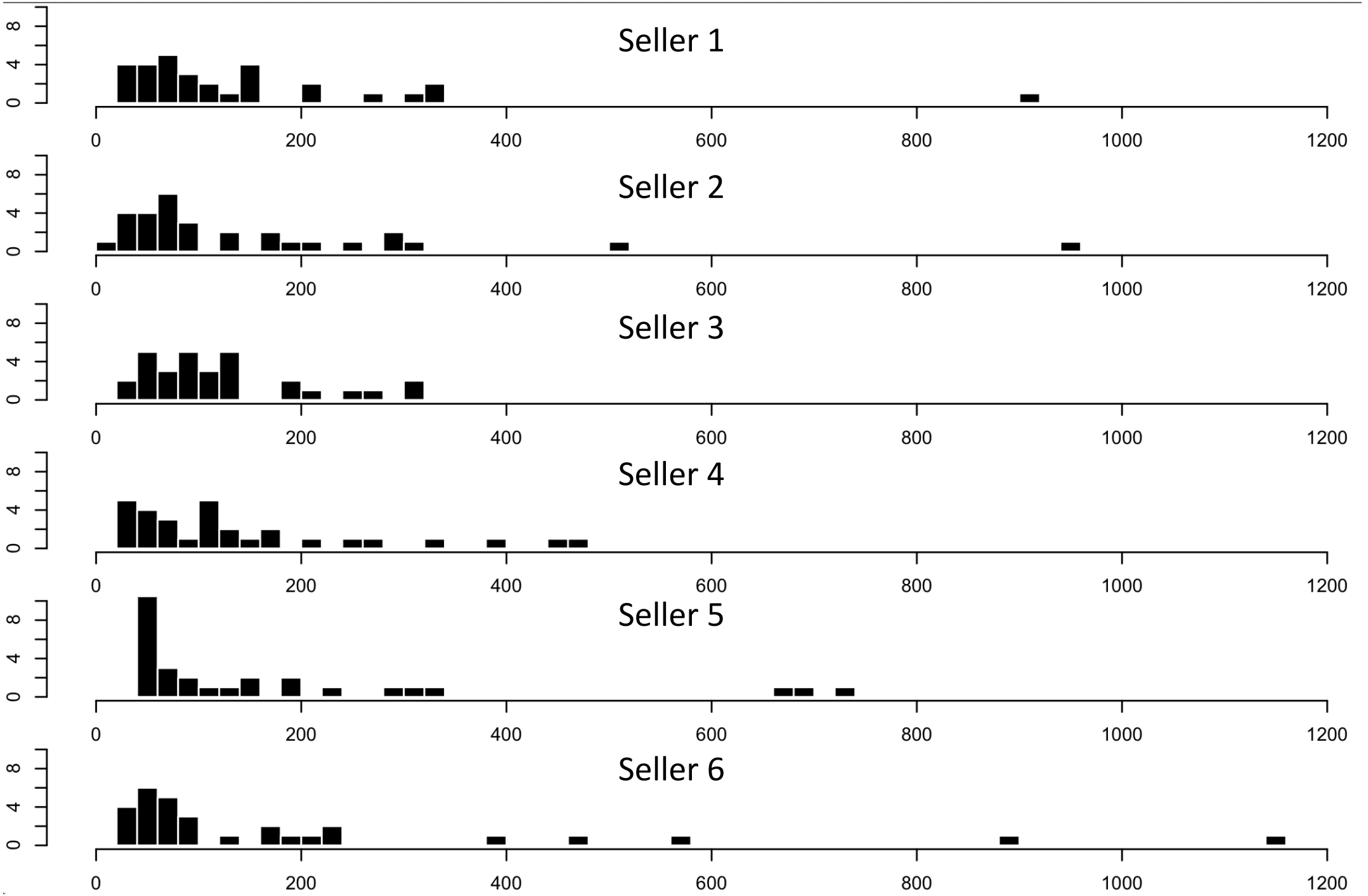
- Selling encyclopedias



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

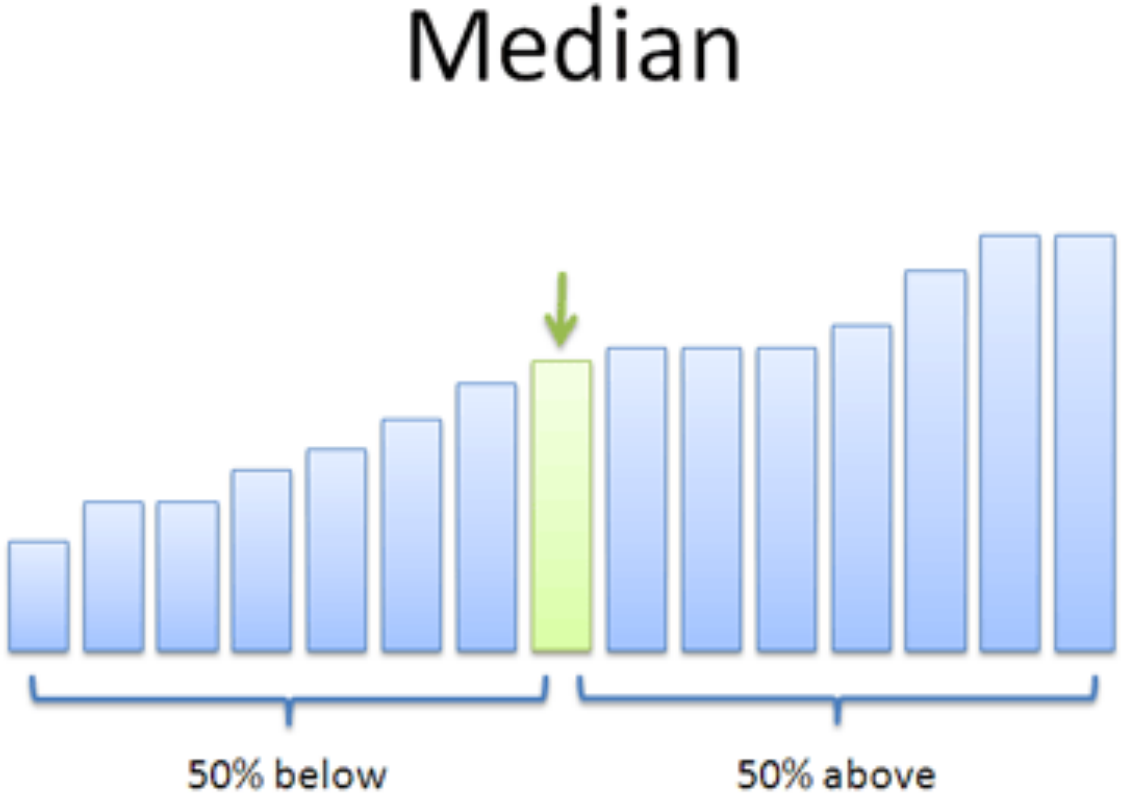

Seller 1
Seller 2
Seller 3
Seller 4
Seller 5
Seller 6
0
4
8
0
200
400
600
800
1000
1200

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

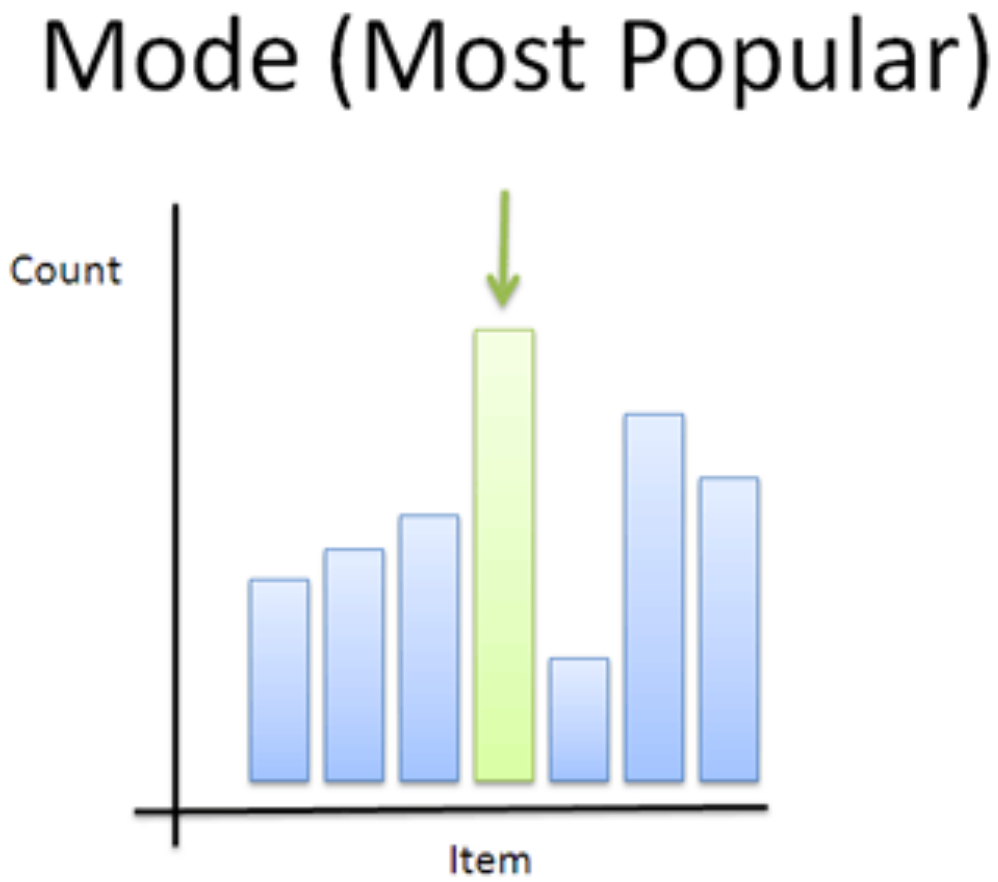
From Kalid Azad

CENTRAL TENDENCY



From Kalid Azad

CENTRAL TENDENCY



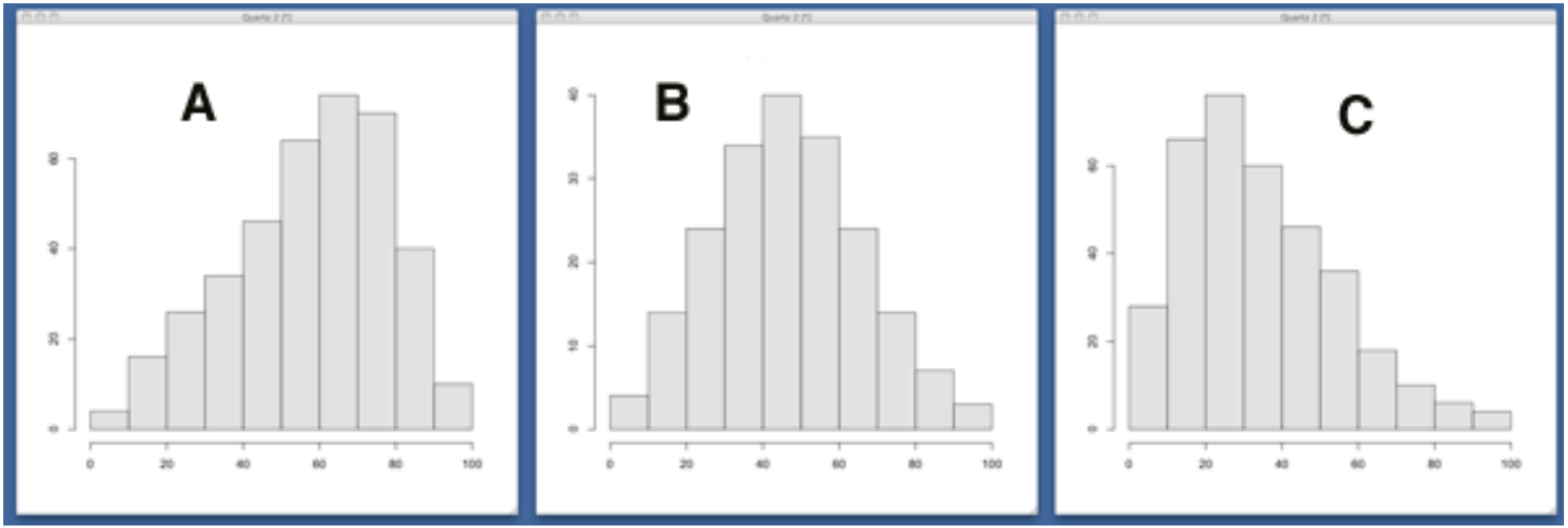
From Kalid Azad

CENTRAL TENDENCY

negative skew

symmetric

positive skew

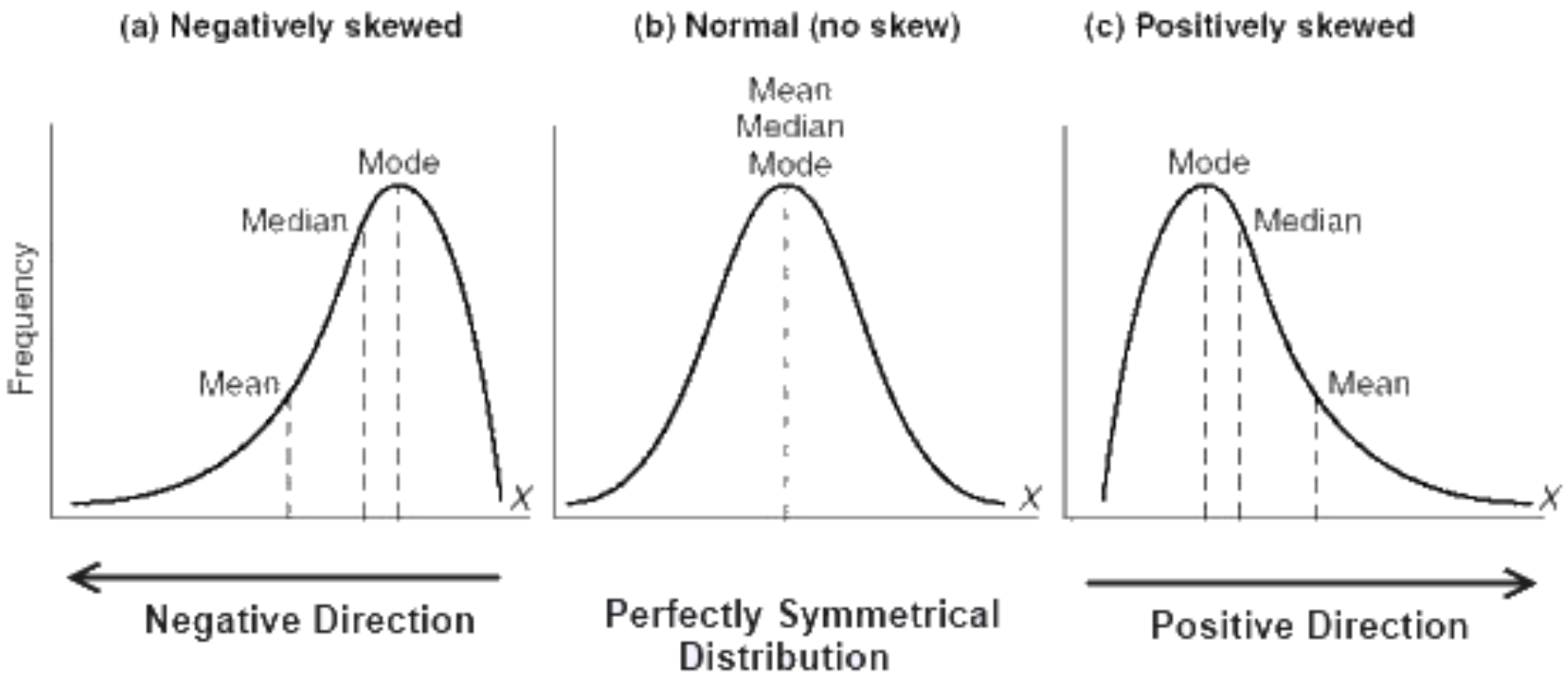


From Shreya Sethi

CENTRAL TENDENCY

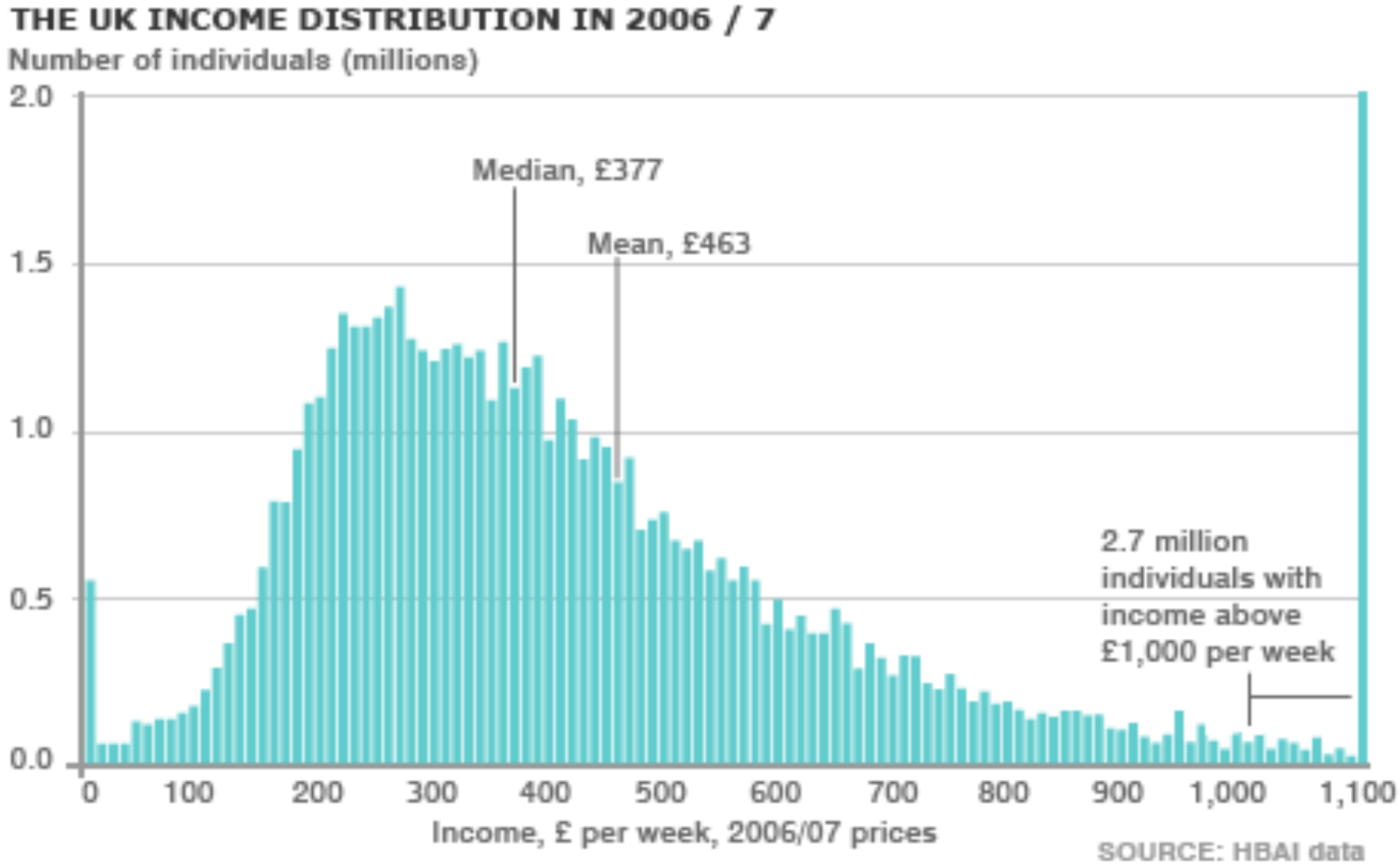
CENTRAL TENDENCY

CENTRAL TENDENCY



From Shreya Sethi

CENTRAL TENDENCY



From Michael Blastland

DISPERSION

Standard Deviation

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

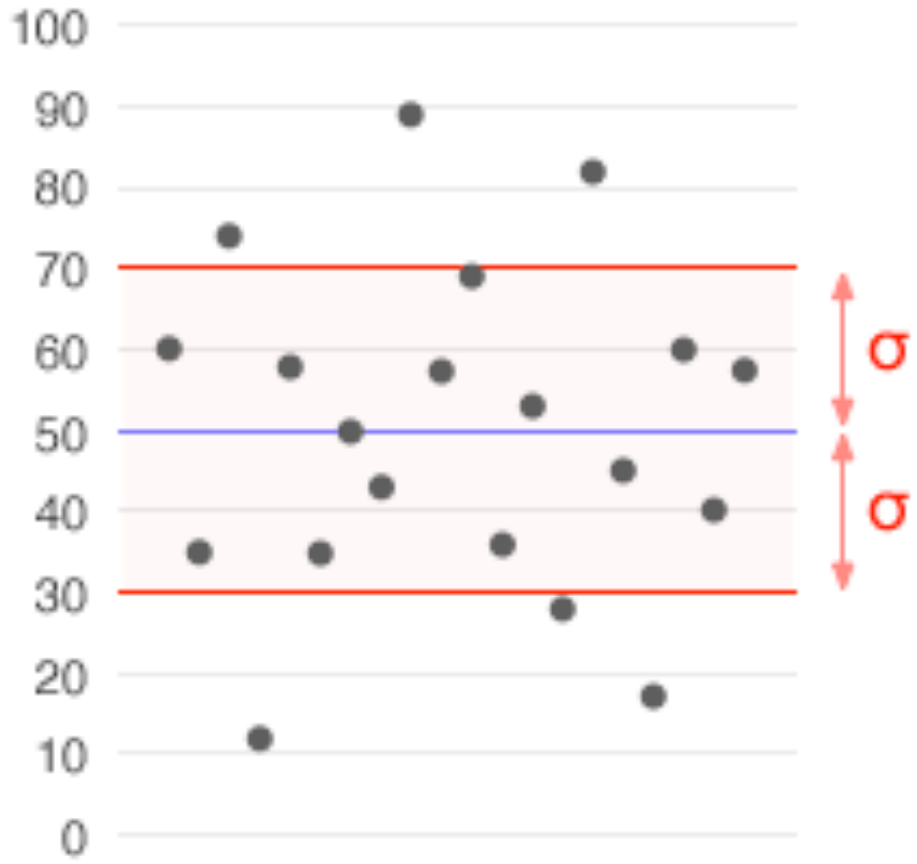
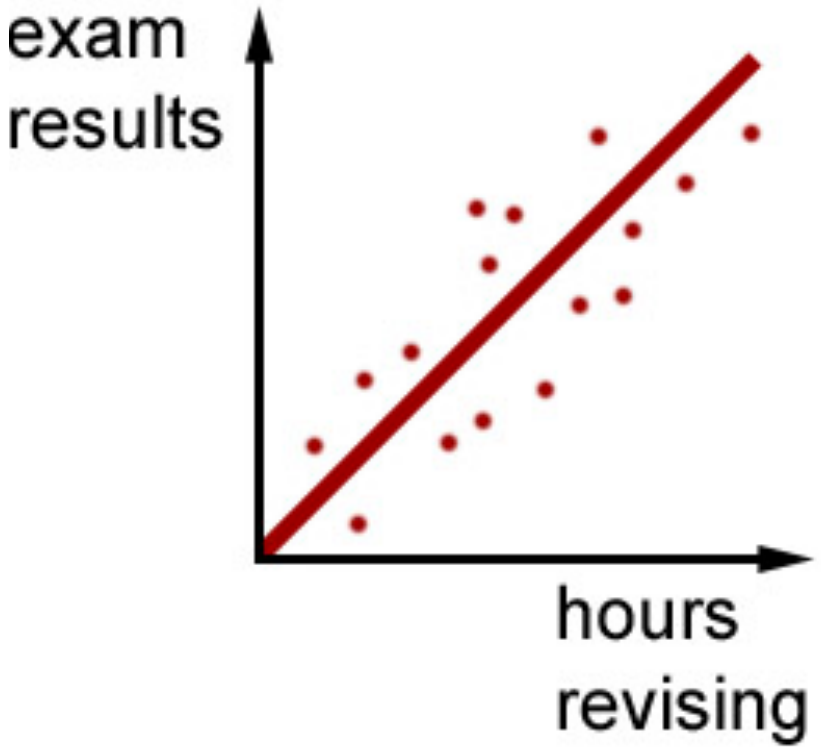


Image from Wikipedia

DEPENDENCE

- Correlation



POSITIVE CORRELATION

- people who do more revision get higher exam results.

Source unknown

DEPENDENCE

- Correlation

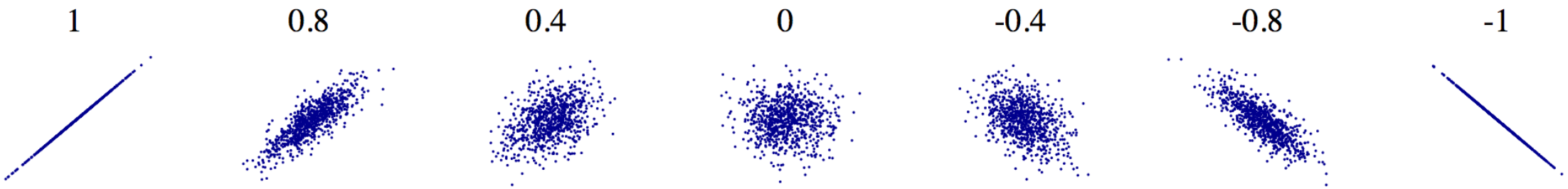
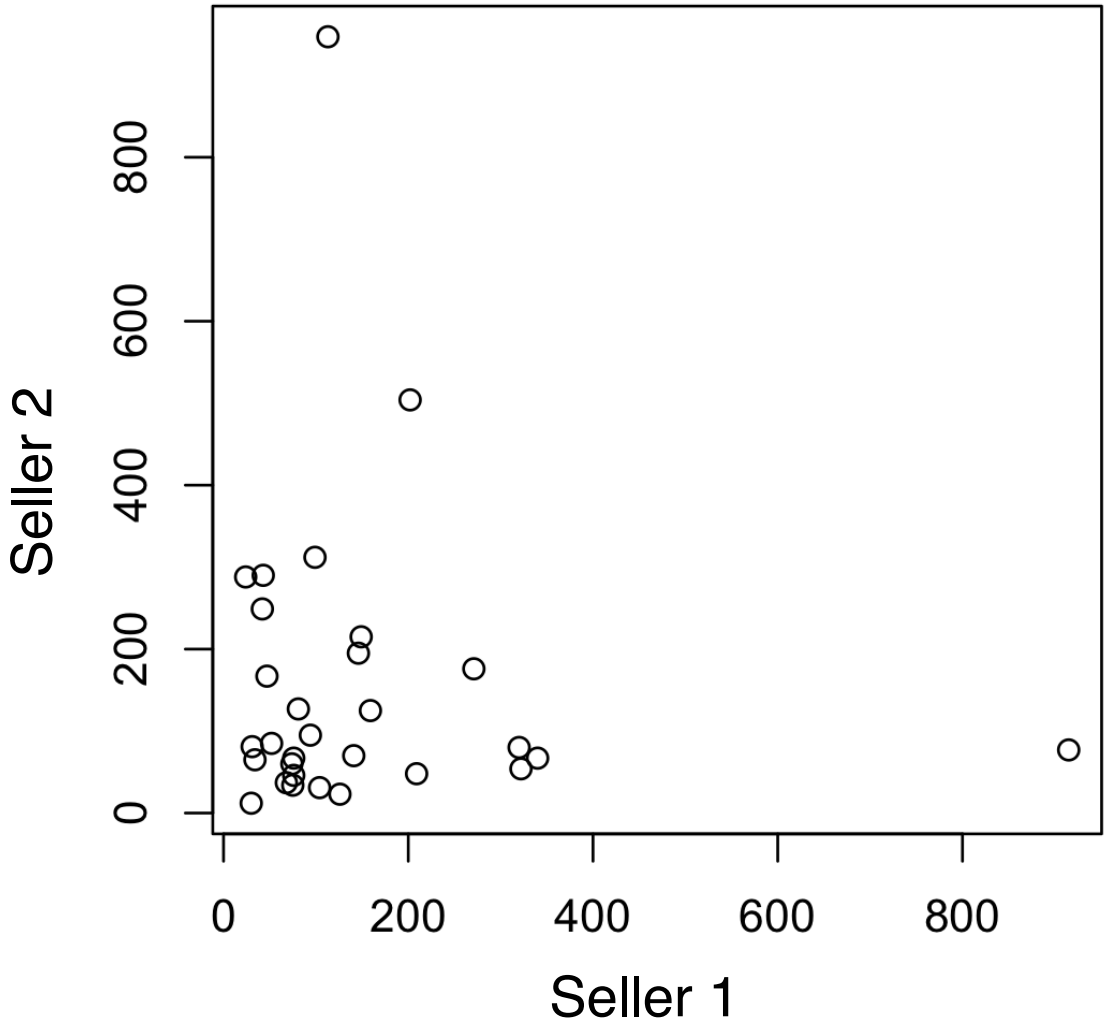


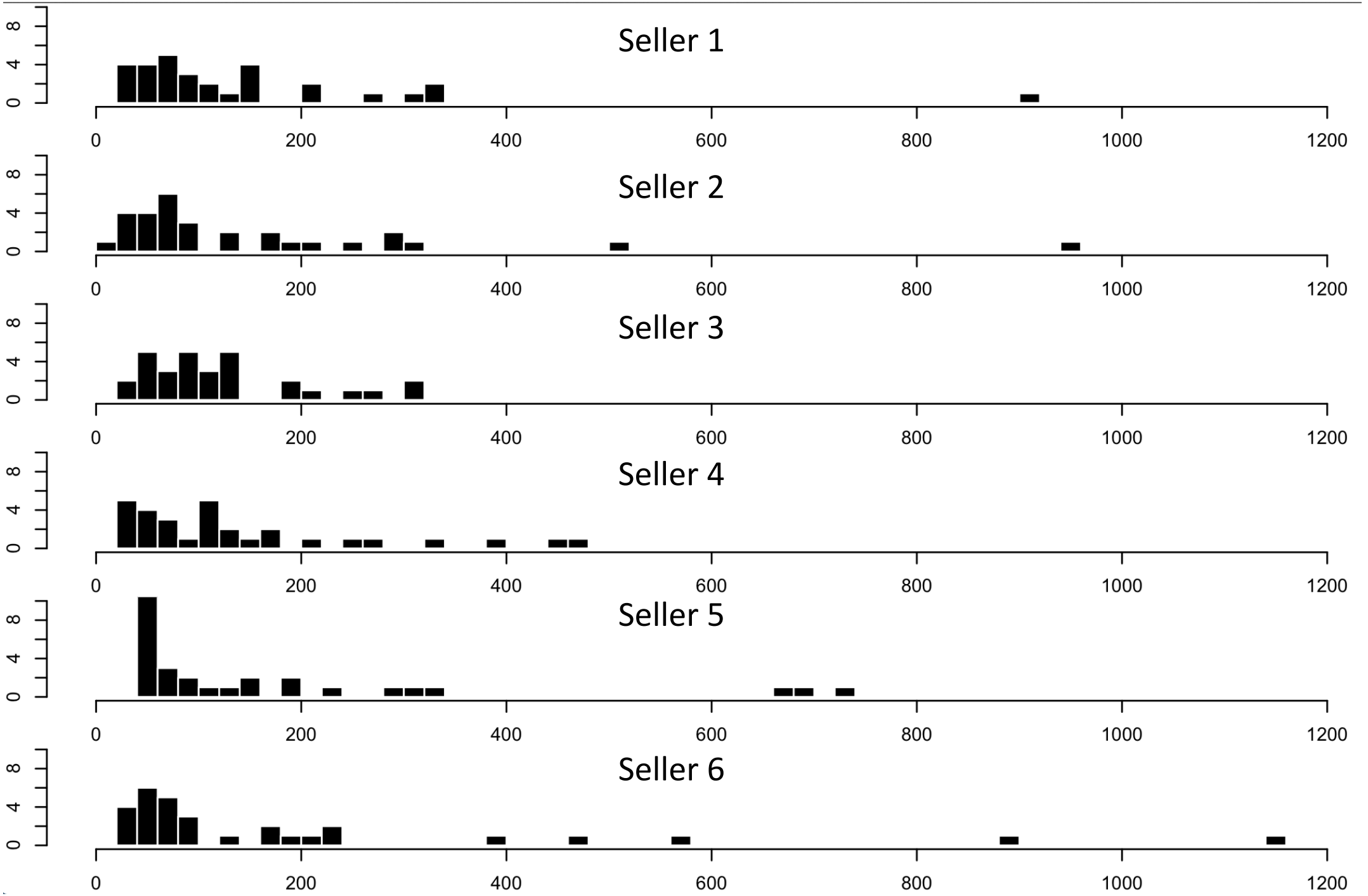
Image from Wikipedia

DEPENDENCE

- Correlation

r = -0.08



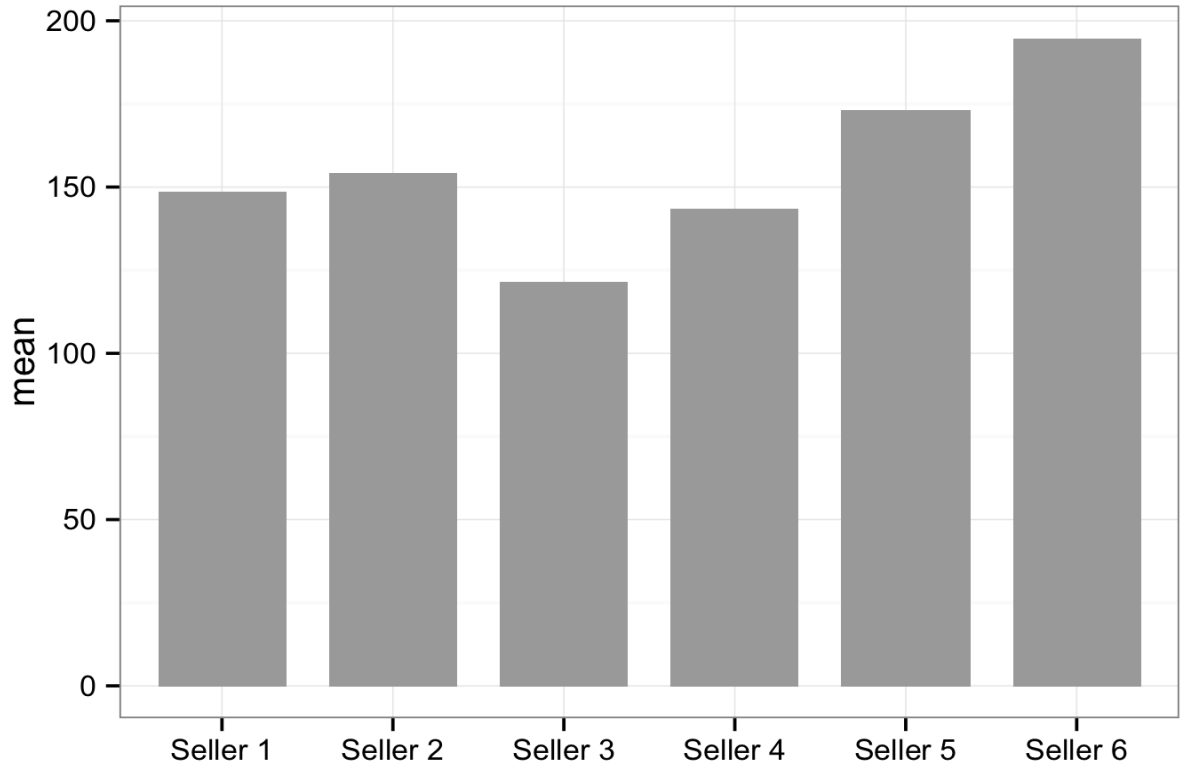

Seller 1
Seller 2
Seller 3
Seller 4
Seller 5
Seller 6
0
4
8
0
200
400
600
800
1000
1200

Average Sales

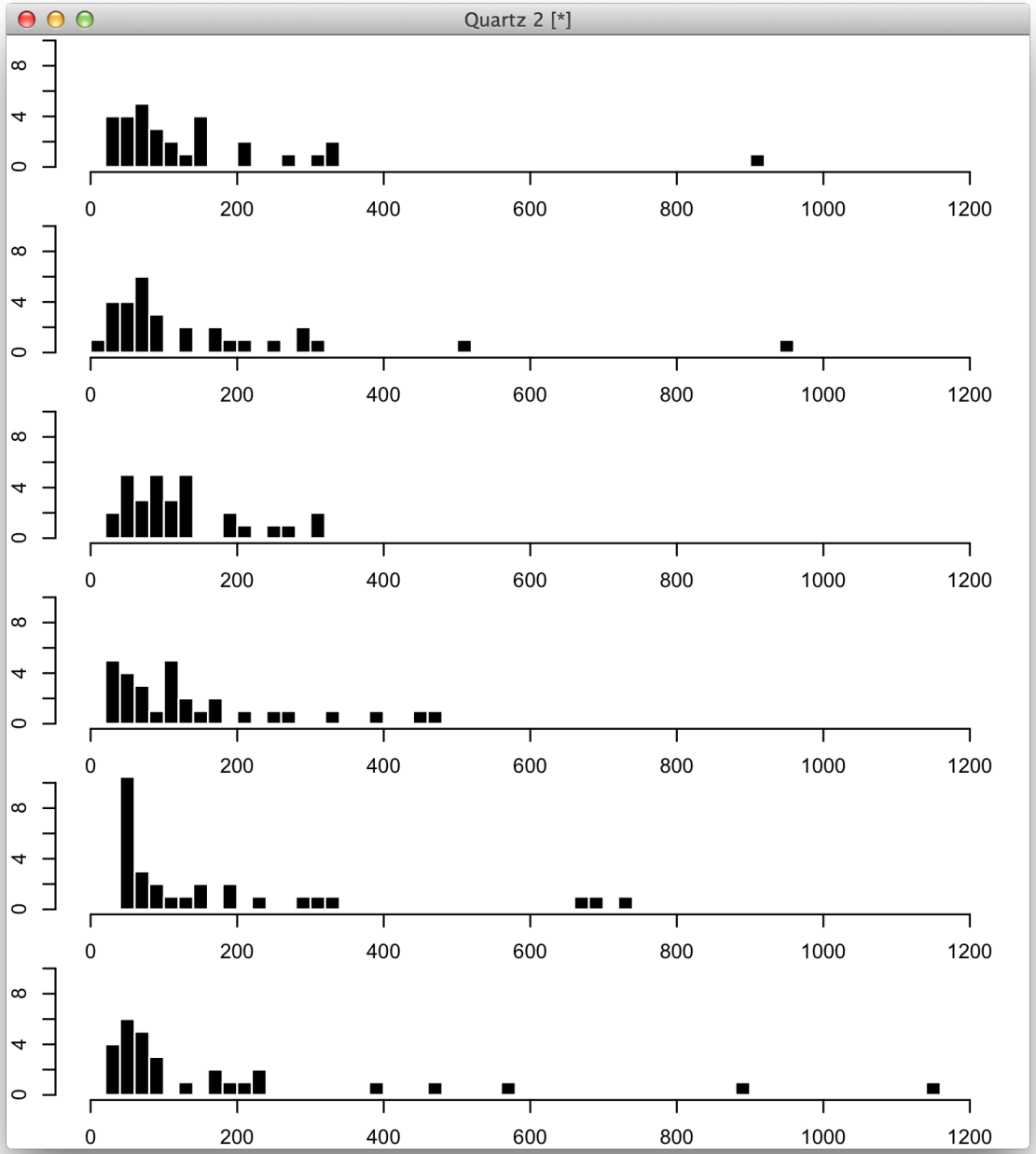
Seller 1	Seller 2	Seller 3	Seller 4	Seller 5	Seller 6
€149	€154	€122	€143	€173	€195

Average Sales

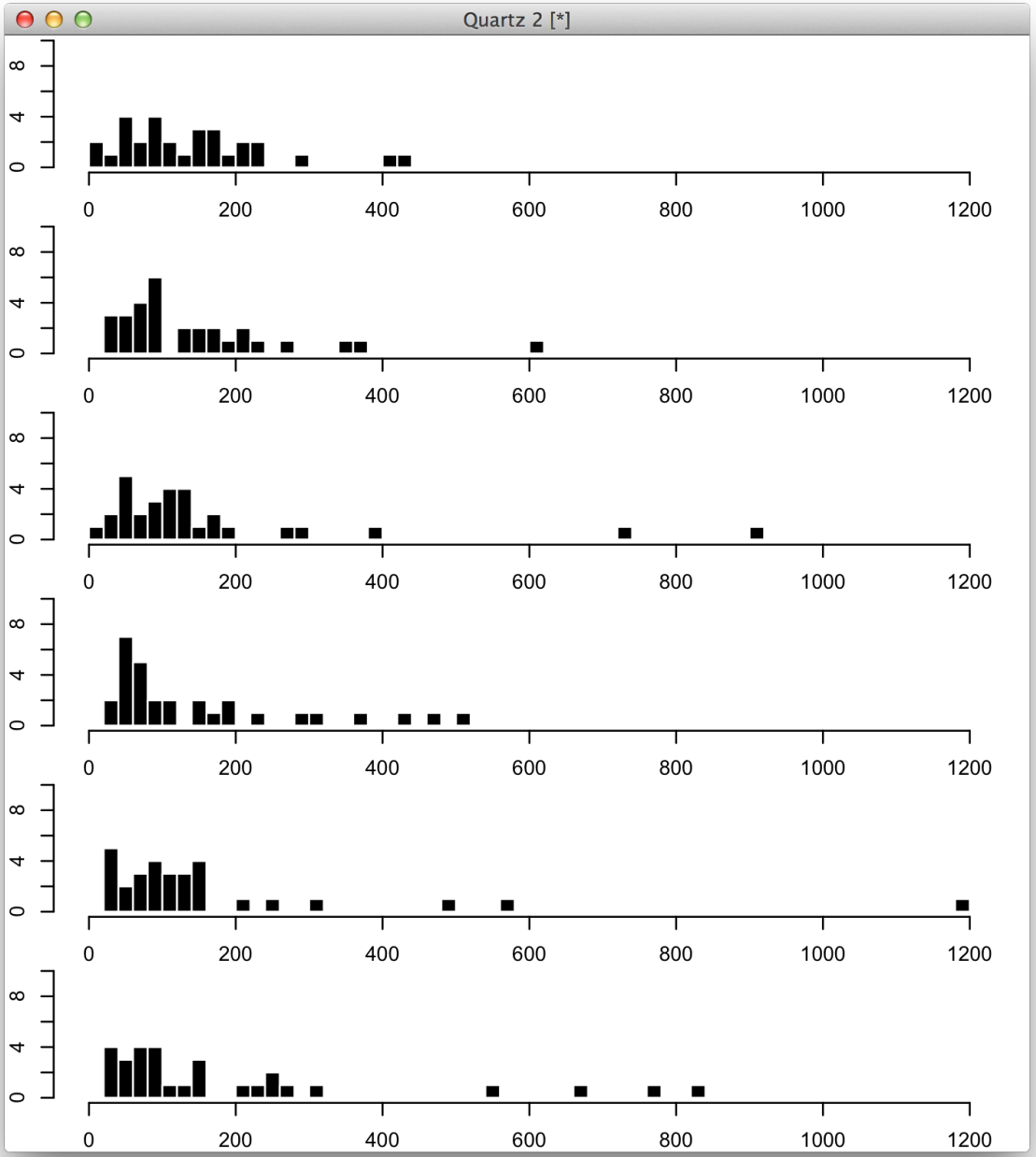
Seller 1	Seller 2	Seller 3	Seller 4	Seller 5	Seller 6
€149	€154	€122	€143	€173	€195



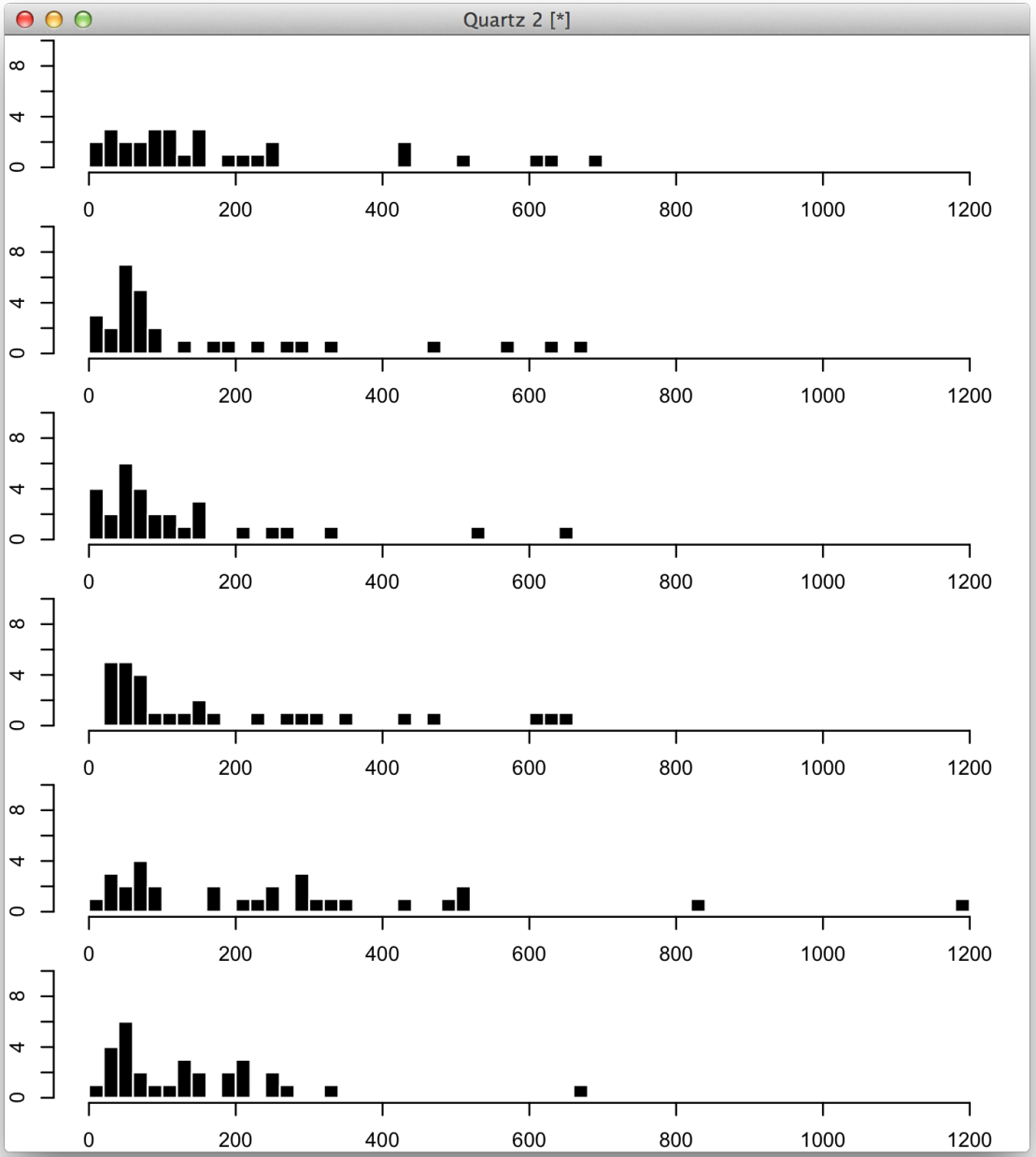
COLLECTING MORE DATA



September 2014

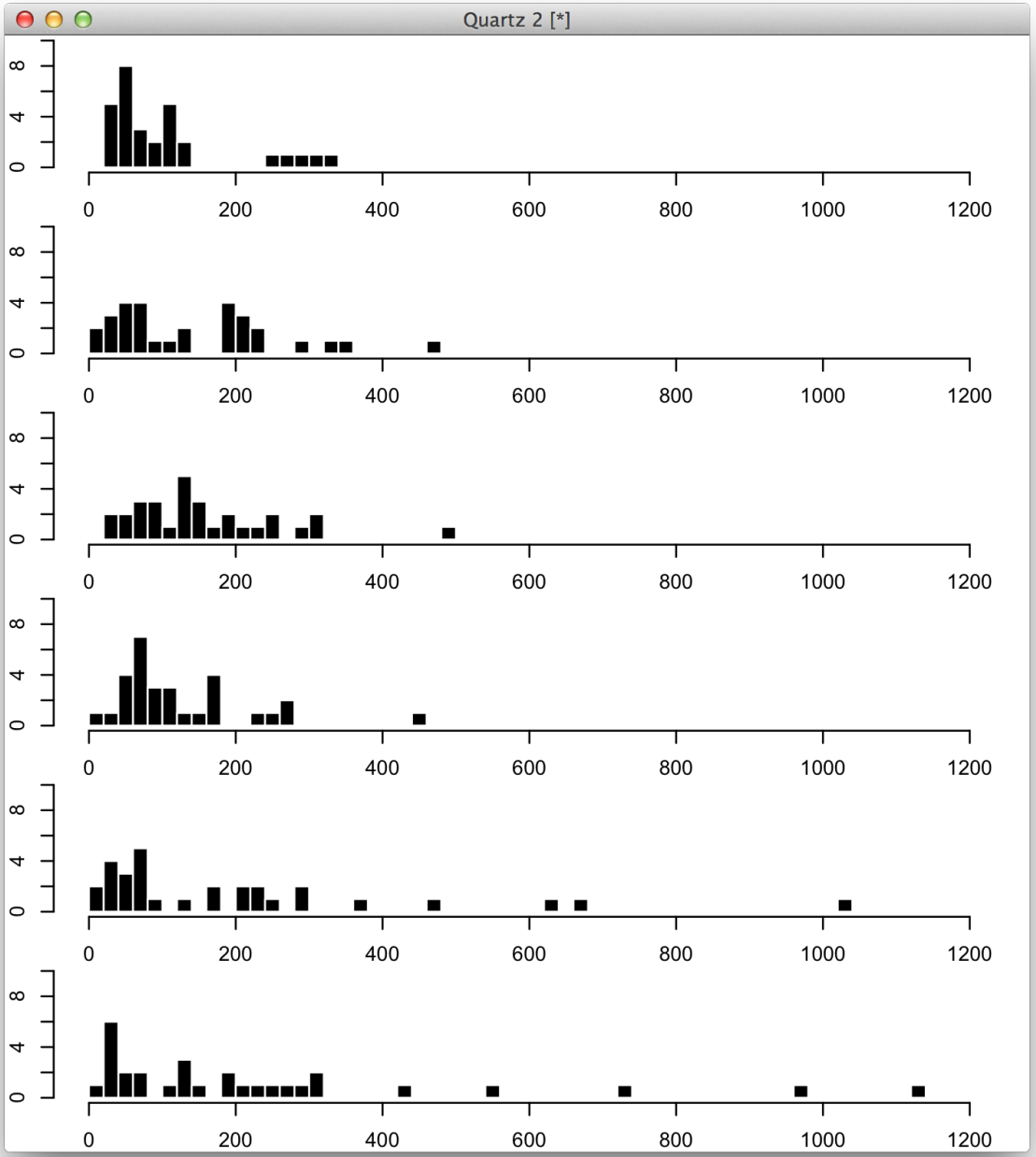


October 2014

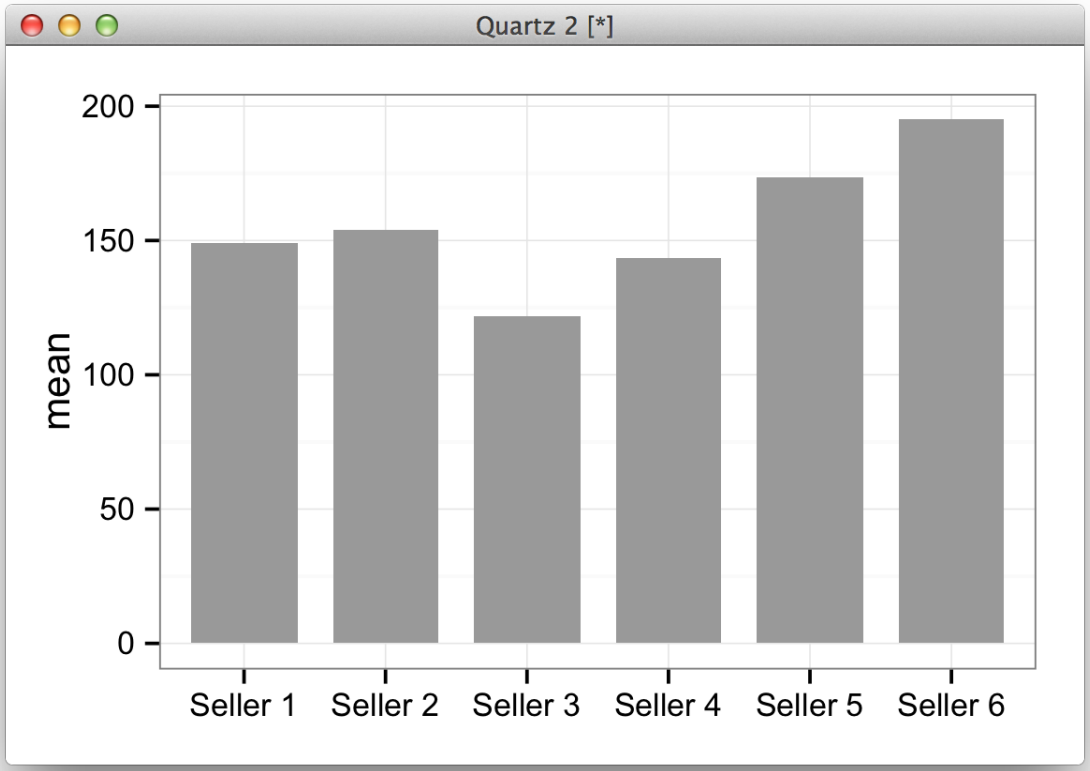


November 2014

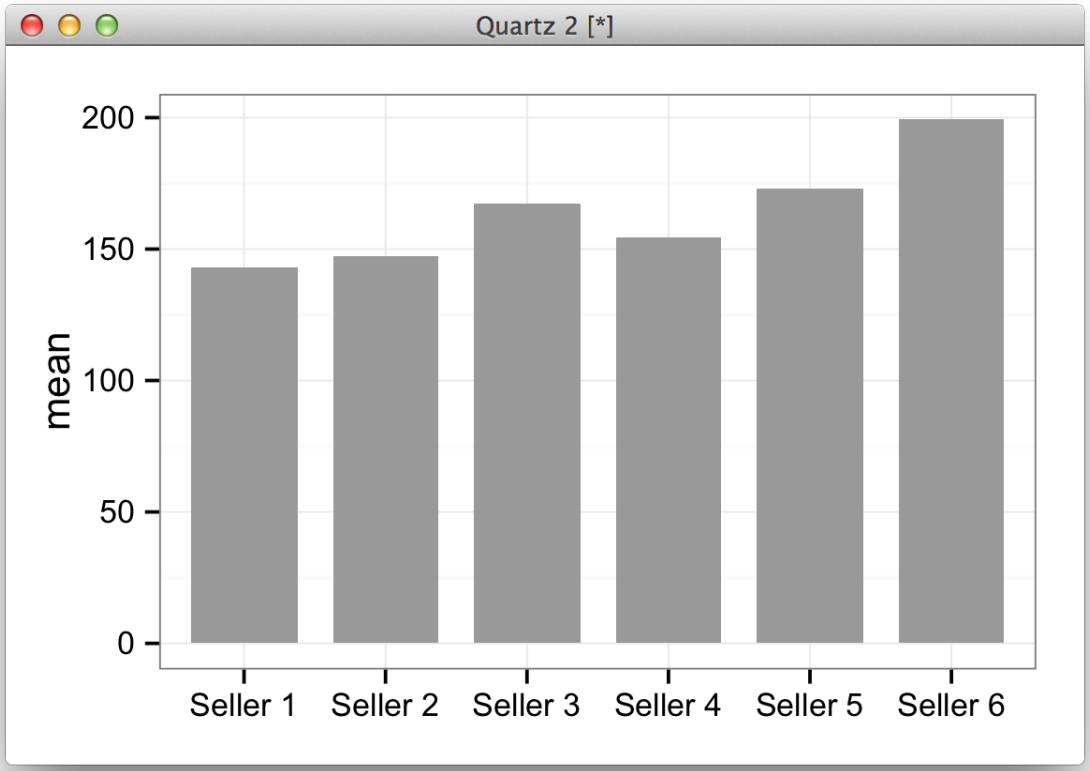
december 2014



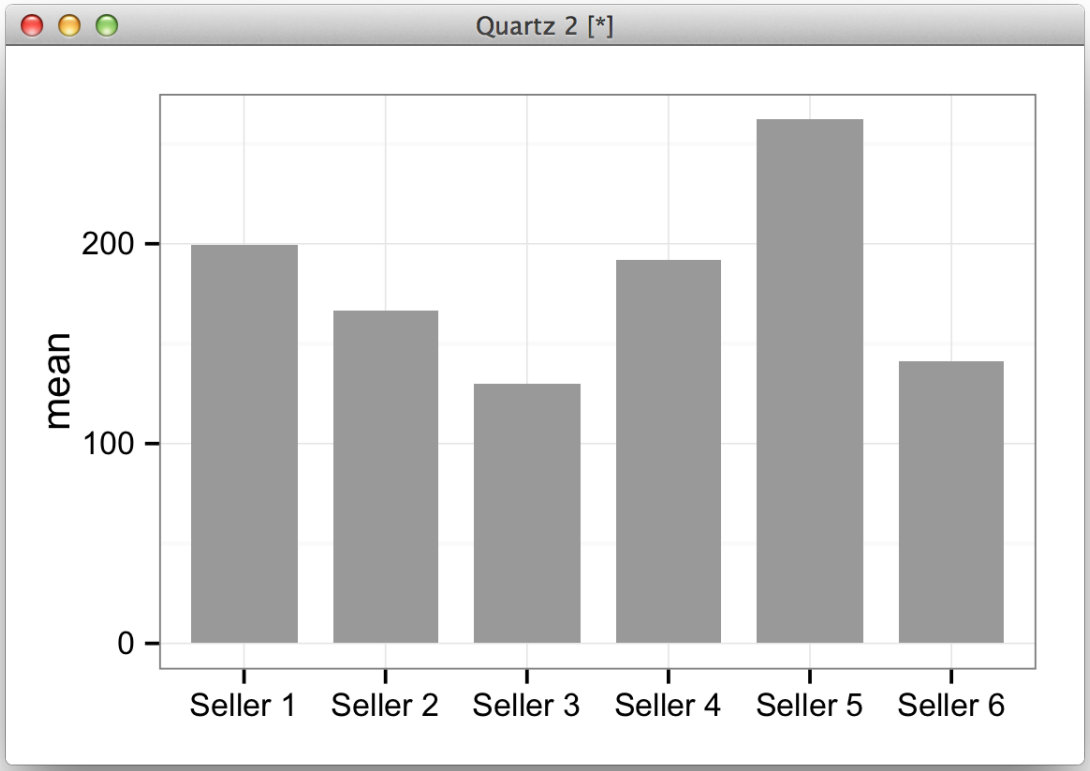
September 2014



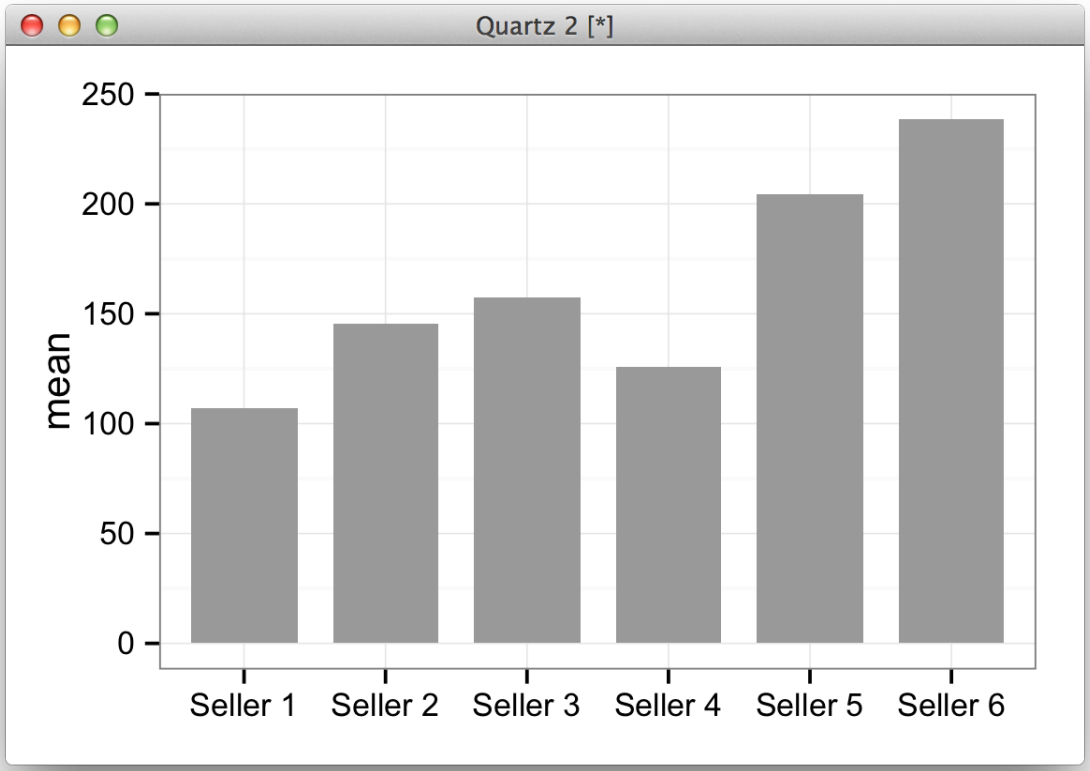
October 2014



November 2014



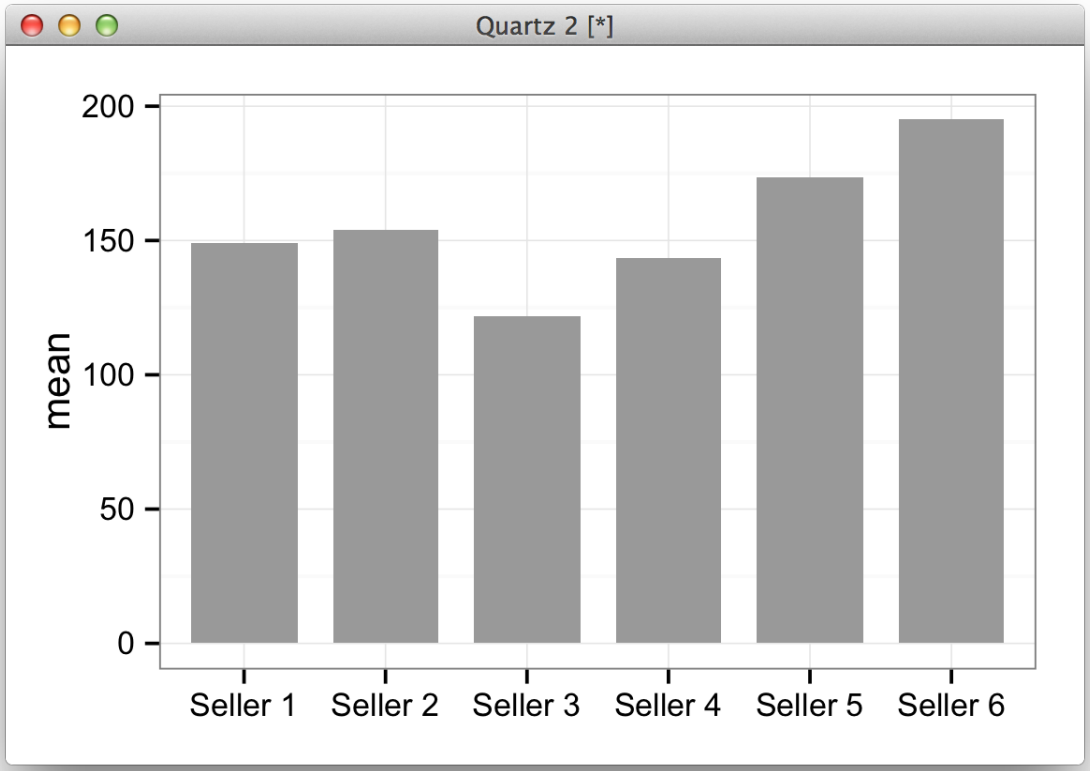
December 2014



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

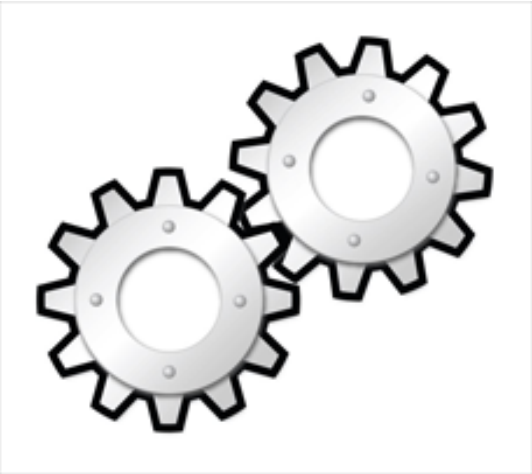
September 2014



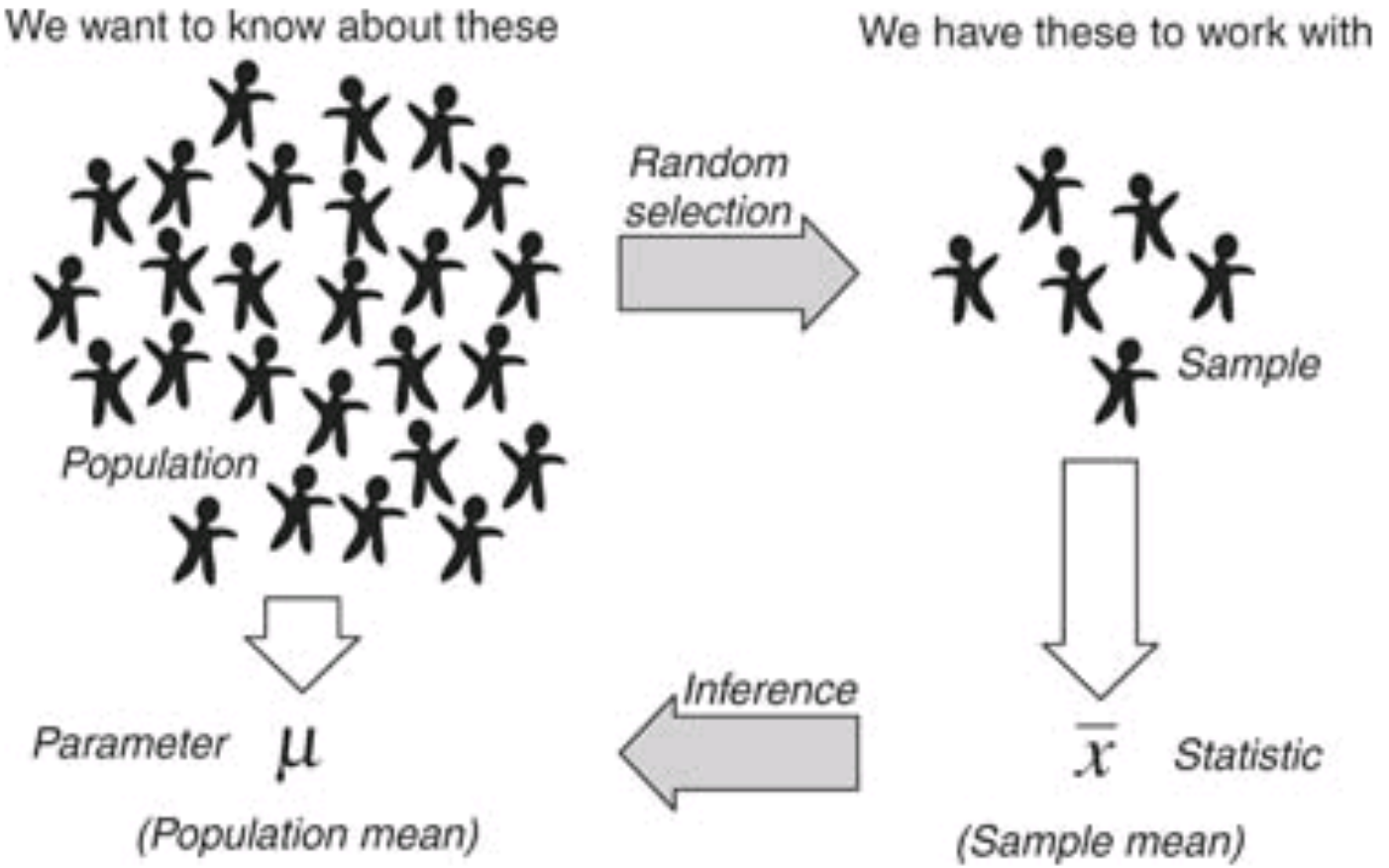
How much can we trust this chart?

STATISTICAL TOOLS

INFERENTIAL STATISTICS



SAMPLING ERROR



SAMPLING ERROR

- Terminology:
  - Population vs. sample
  - Sample **statistic** (mean, median, etc.)
  - Population **parameter** (mean, median, etc.)

SAMPLING ERROR

- 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 ERROR

- 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

SAMPLING ERROR

- 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

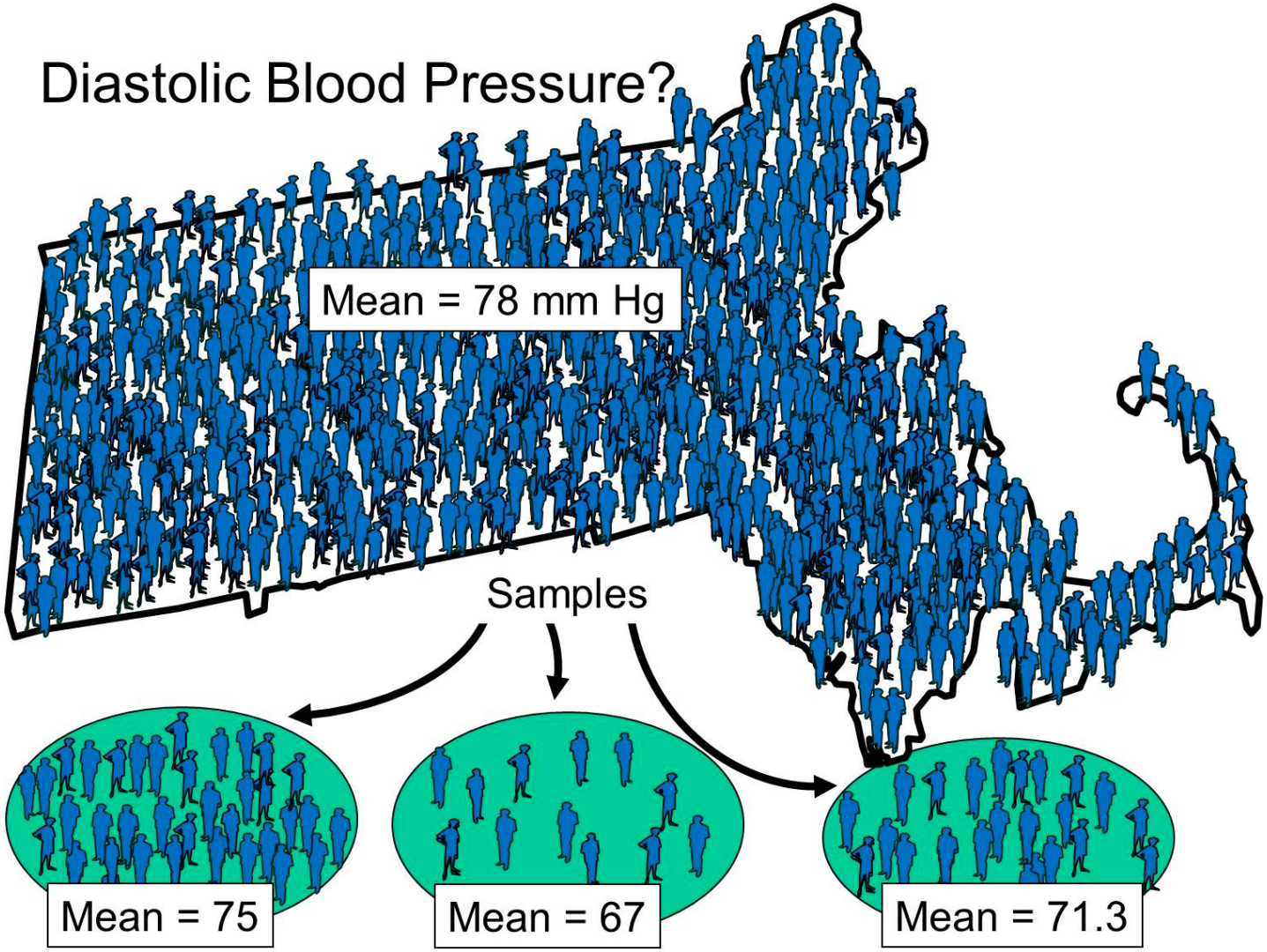
SAMPLING ERROR

- Unit of statistical analysis

Average Sales

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

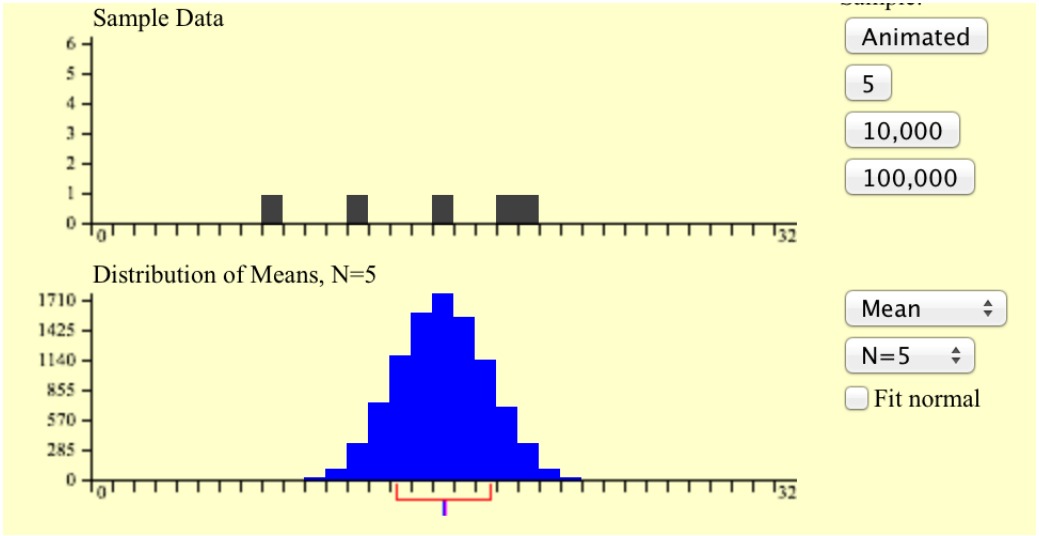
SAMPLING ERROR



From Lisa Sullivan

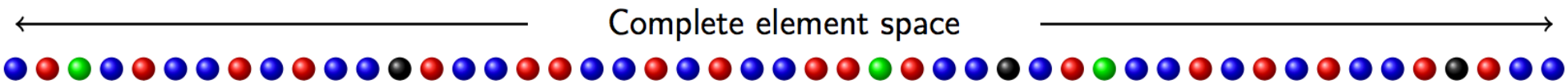
SAMPLING ERROR

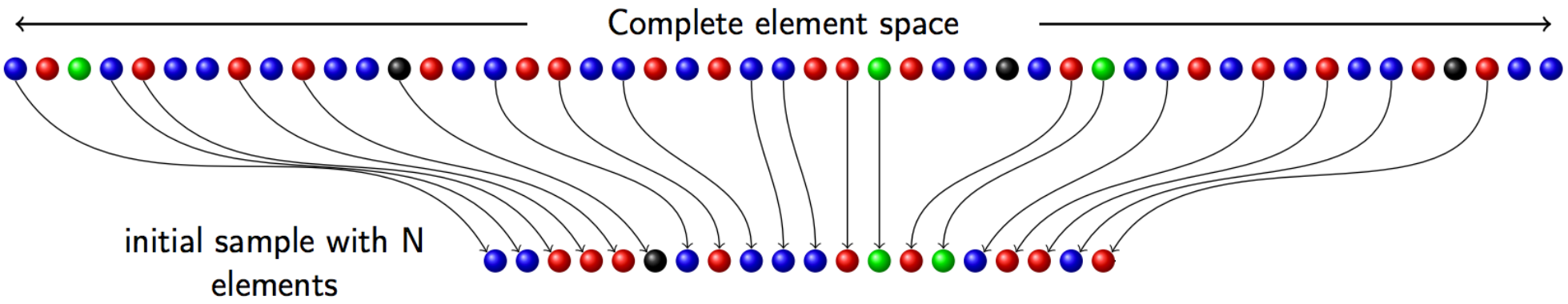
- Sampling distribution of a statistic
  - Demo

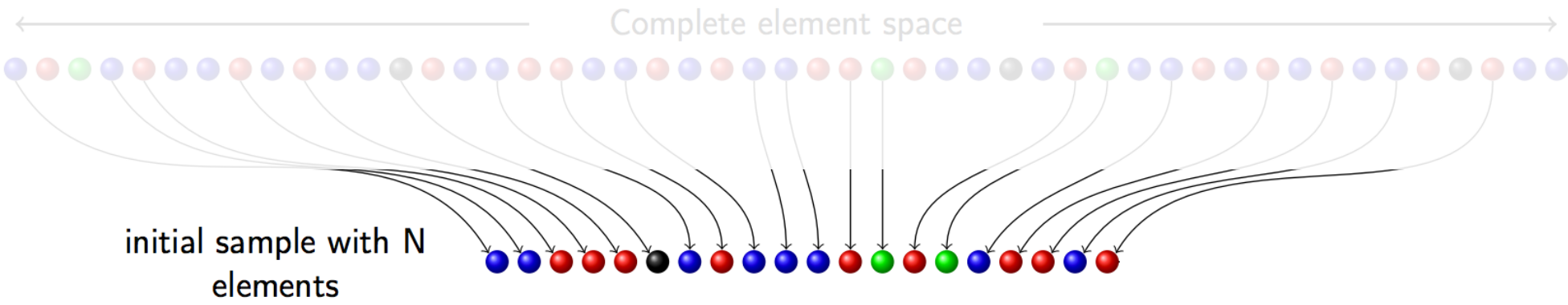


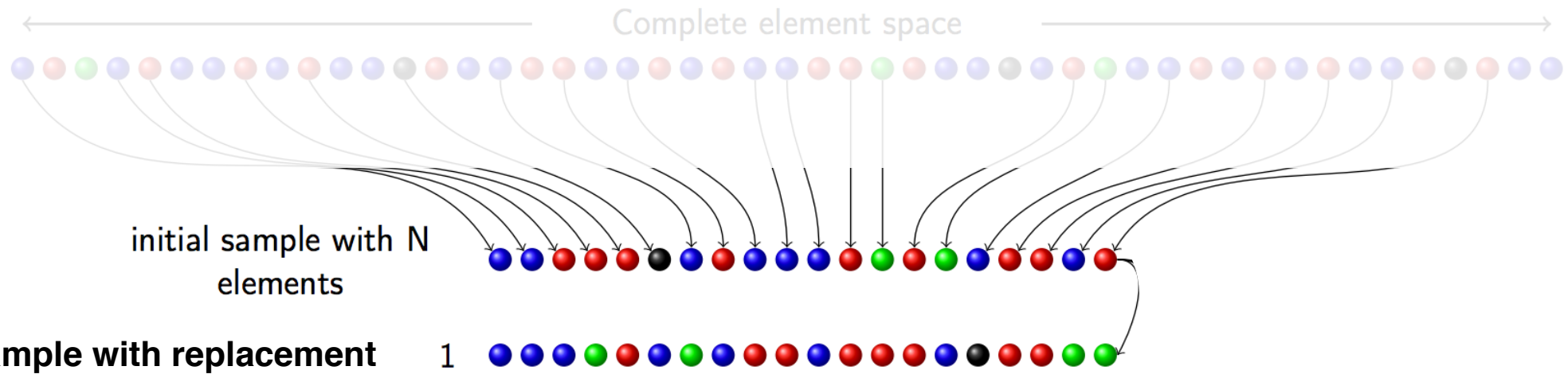
SAMPLING ERROR

- Resampling techniques
  - Bootstrapping

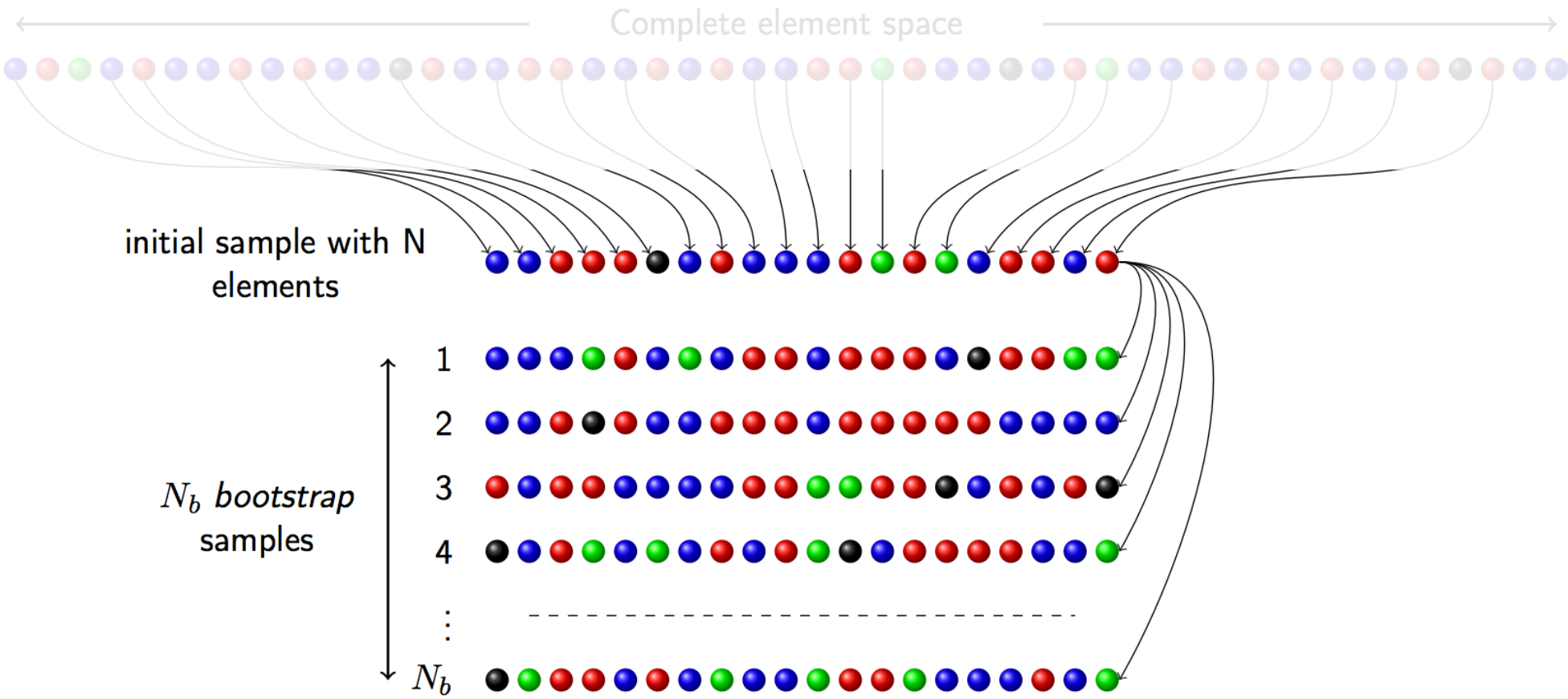


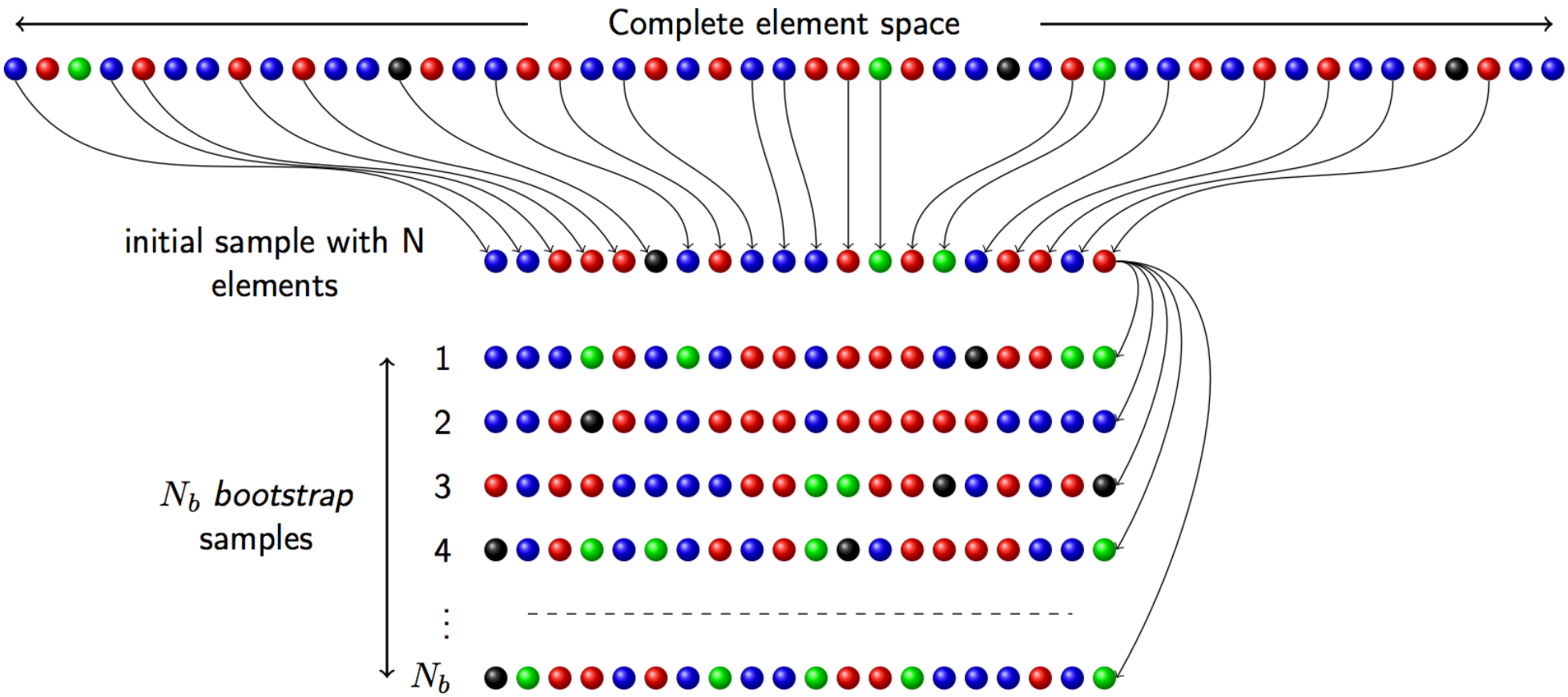






From Germain Salvato-Vallverdu





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

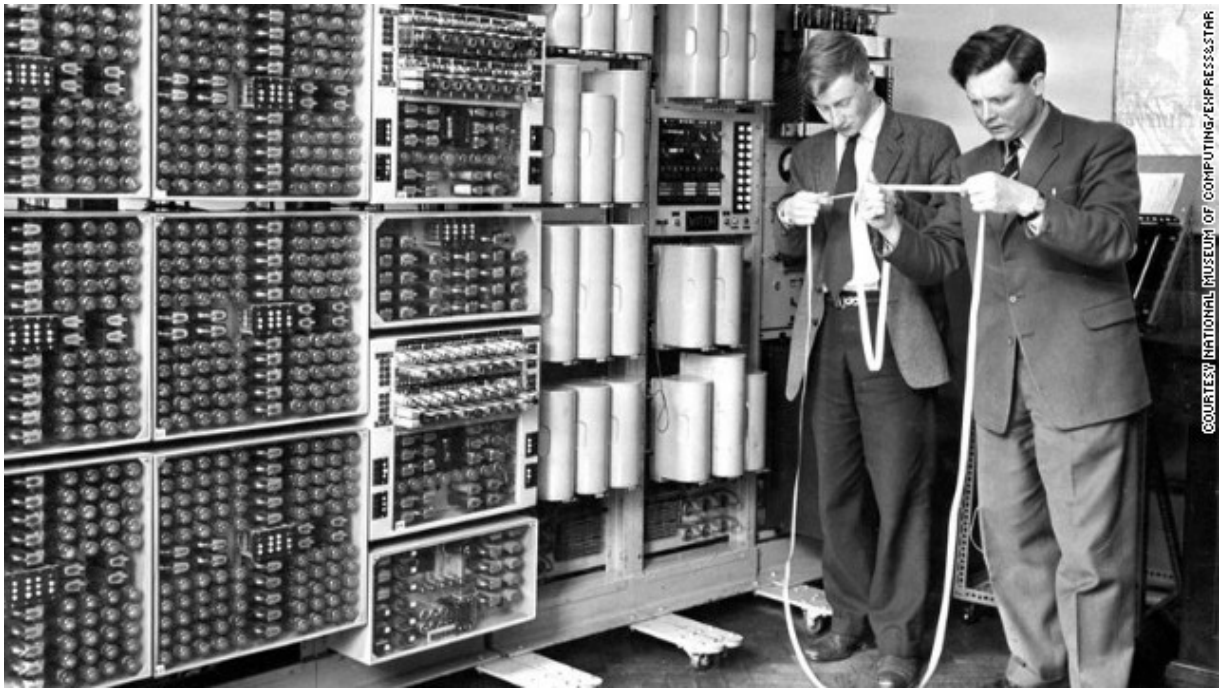
When N tend to infinity, the distribution of average values computed from bootstrap samples is equal to the distribution of average values obtained from ALL samples with N elements which can be constructed from the complete space. Thus the width of the distribution gives an evaluation of the sample quality.

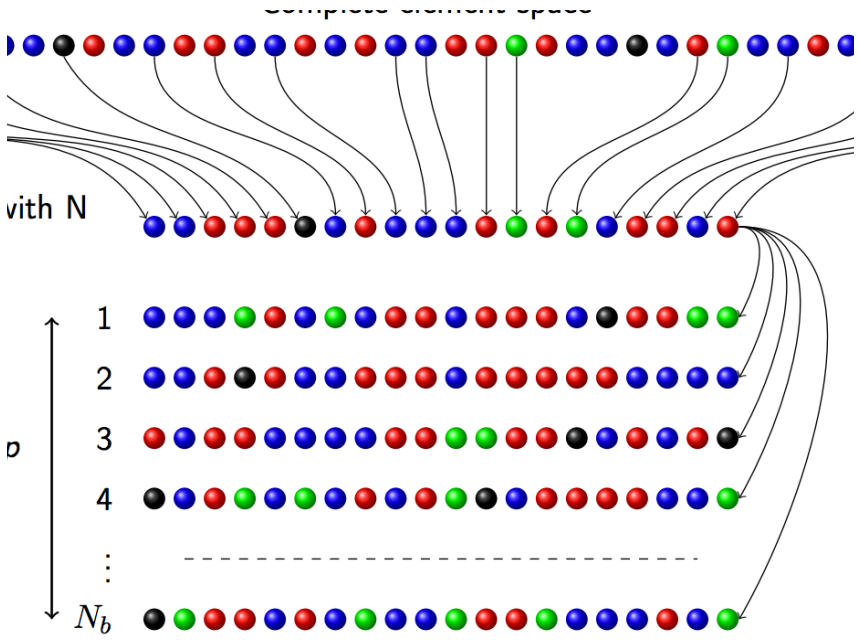
SAMPLING ERROR

- Bootstrapping video

SAMPLING ERROR

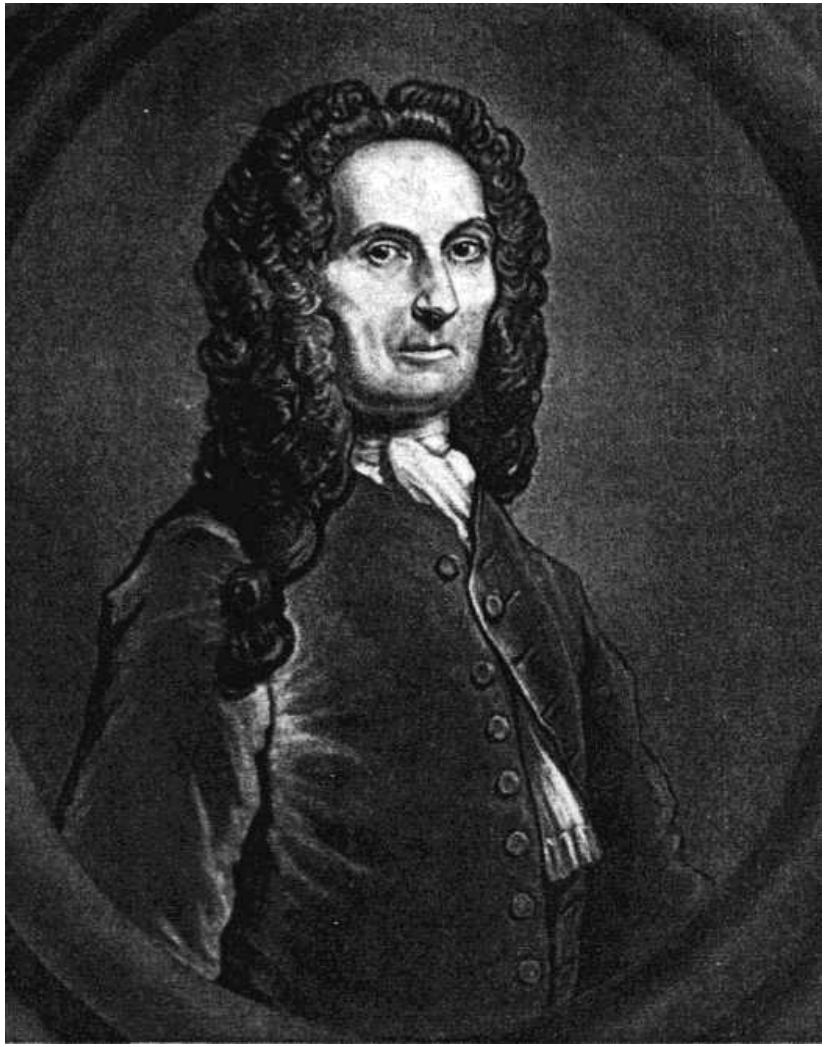
- How did people do before computers?





MORE HISTORY

- Abraham De Moivre
  1667 - 1754



MORE HISTORY

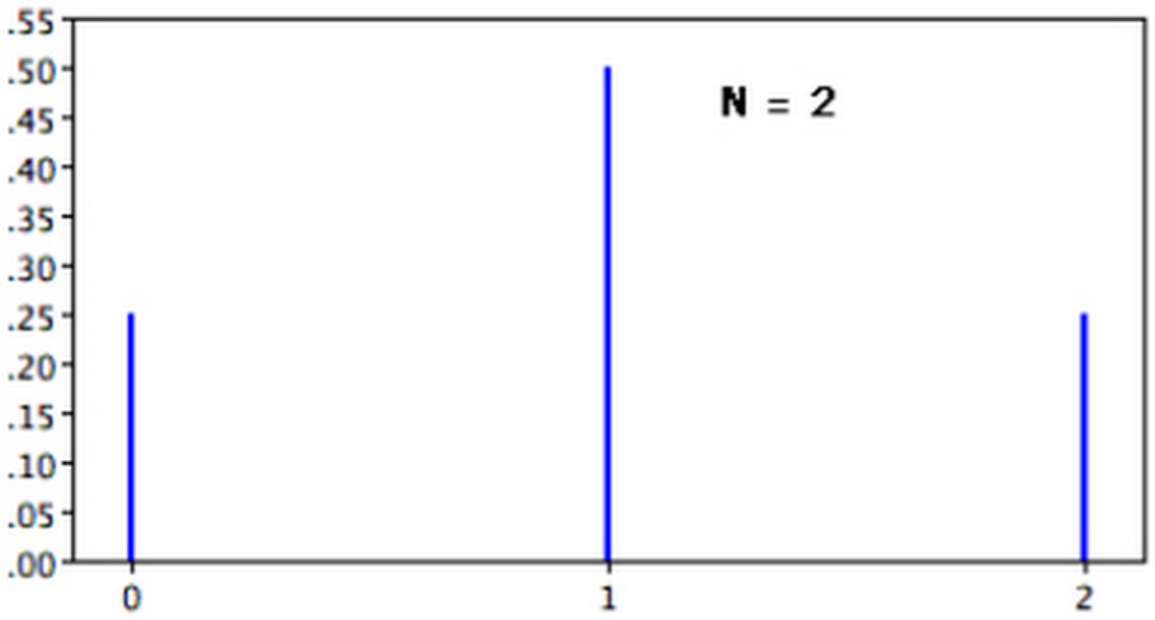
- Abraham De Moivre

  1667 - 1754

MORE HISTORY

- Abraham De Moivre

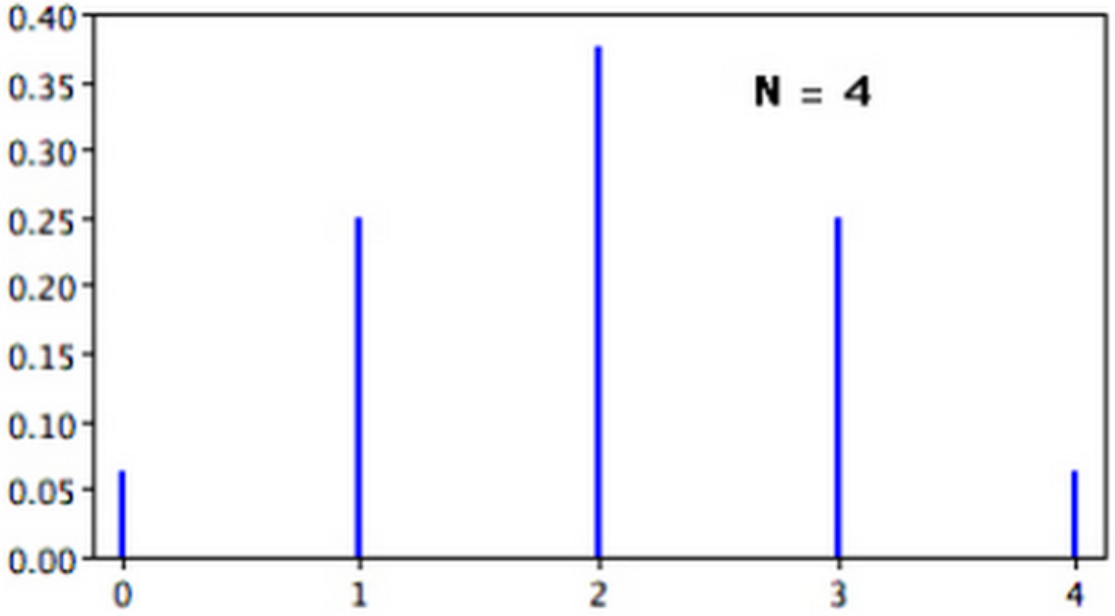
  1667 - 1754



MORE HISTORY

- Abraham De Moivre

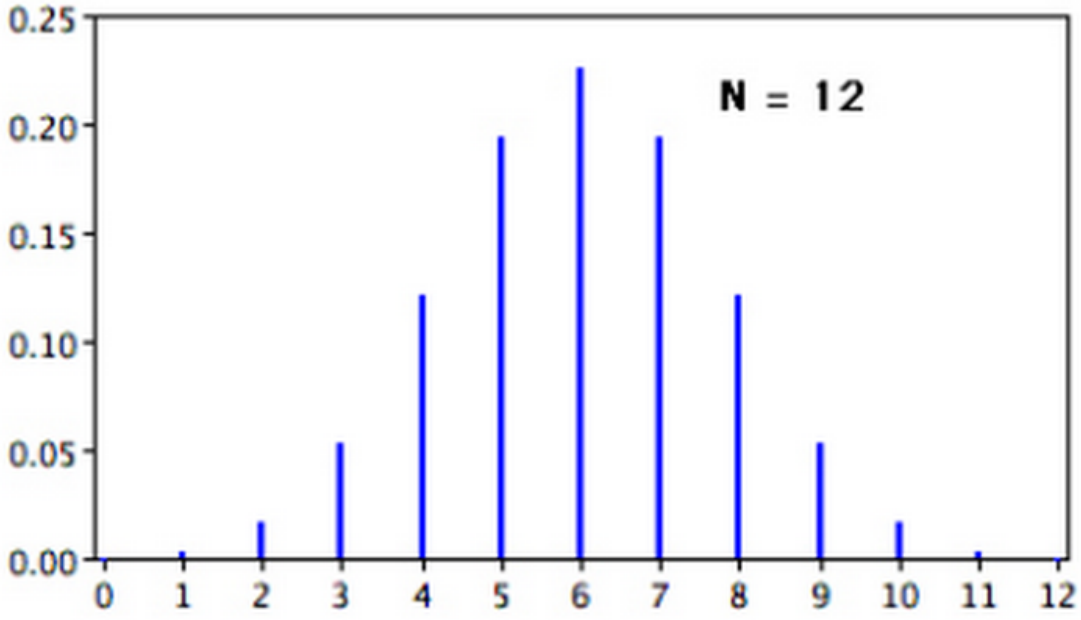
  1667 - 1754



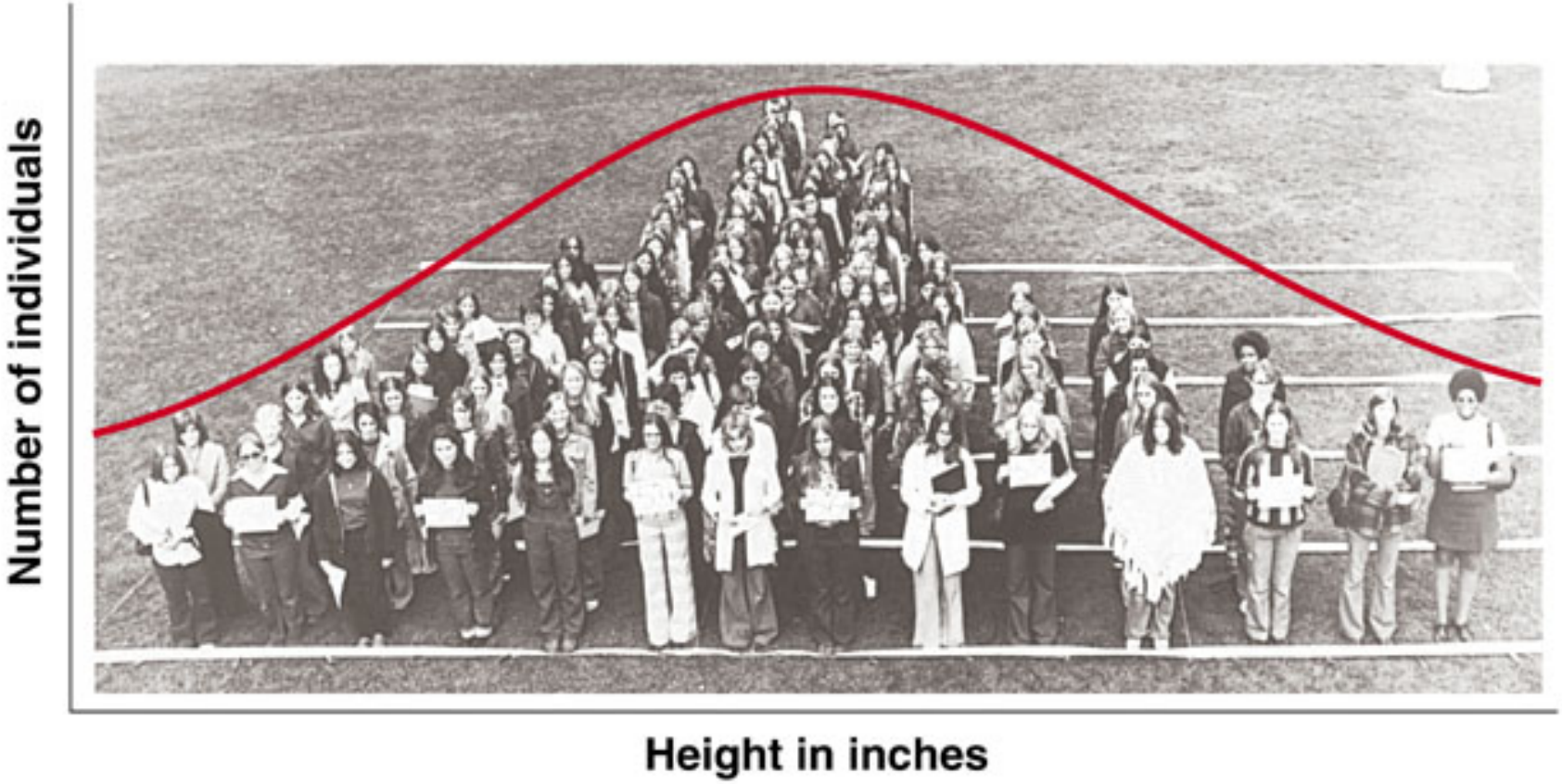
MORE HISTORY

- Abraham De Moivre

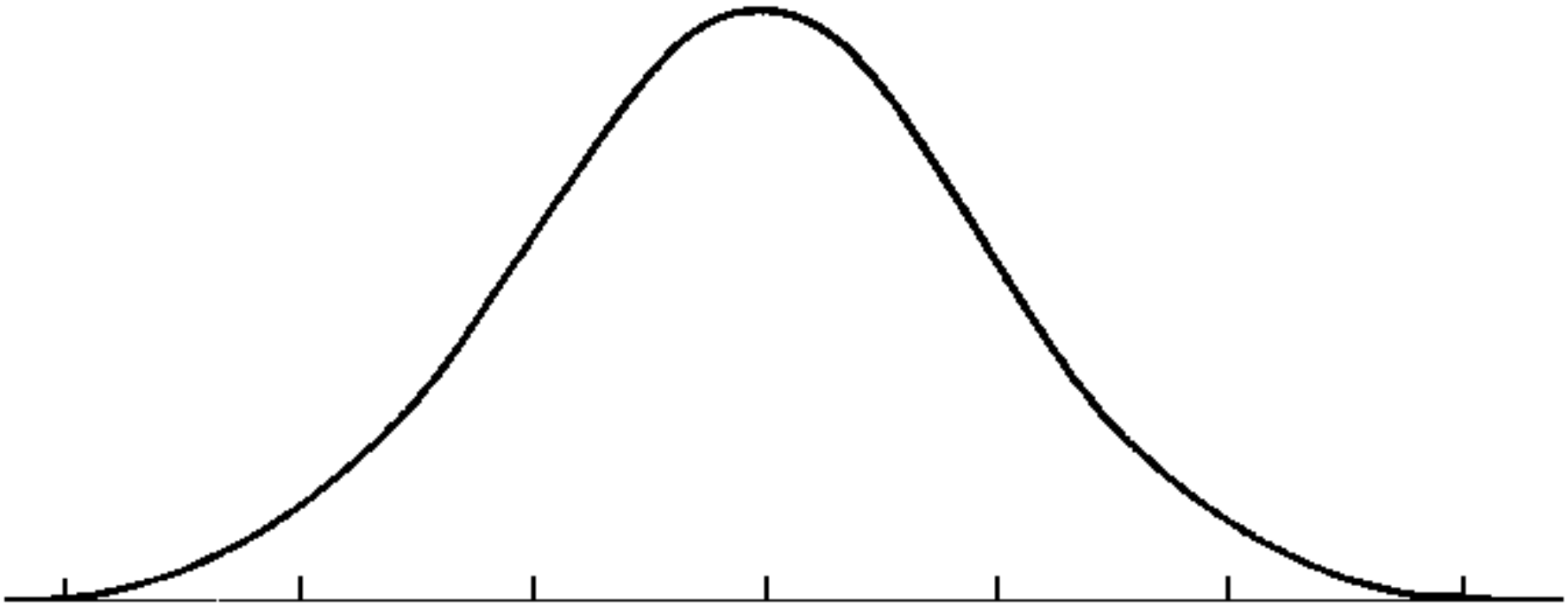
  1667 - 1754



MORE HISTORY



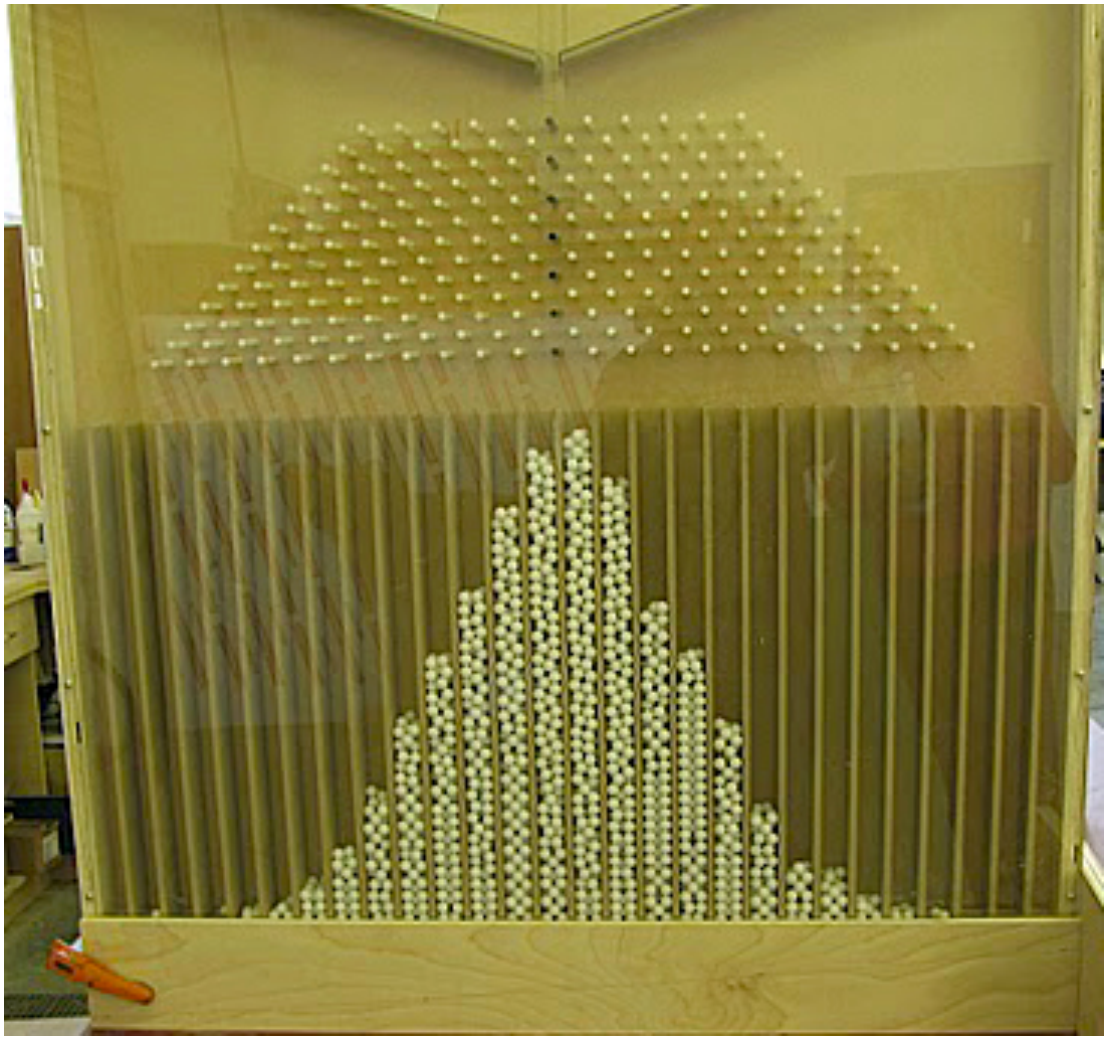
NORMAL DISTRIBUTION



NORMAL DISTRIBUTION

- Sir Francis Galton
  1822 – 1911

  Bean Machine
  or Galton Board:



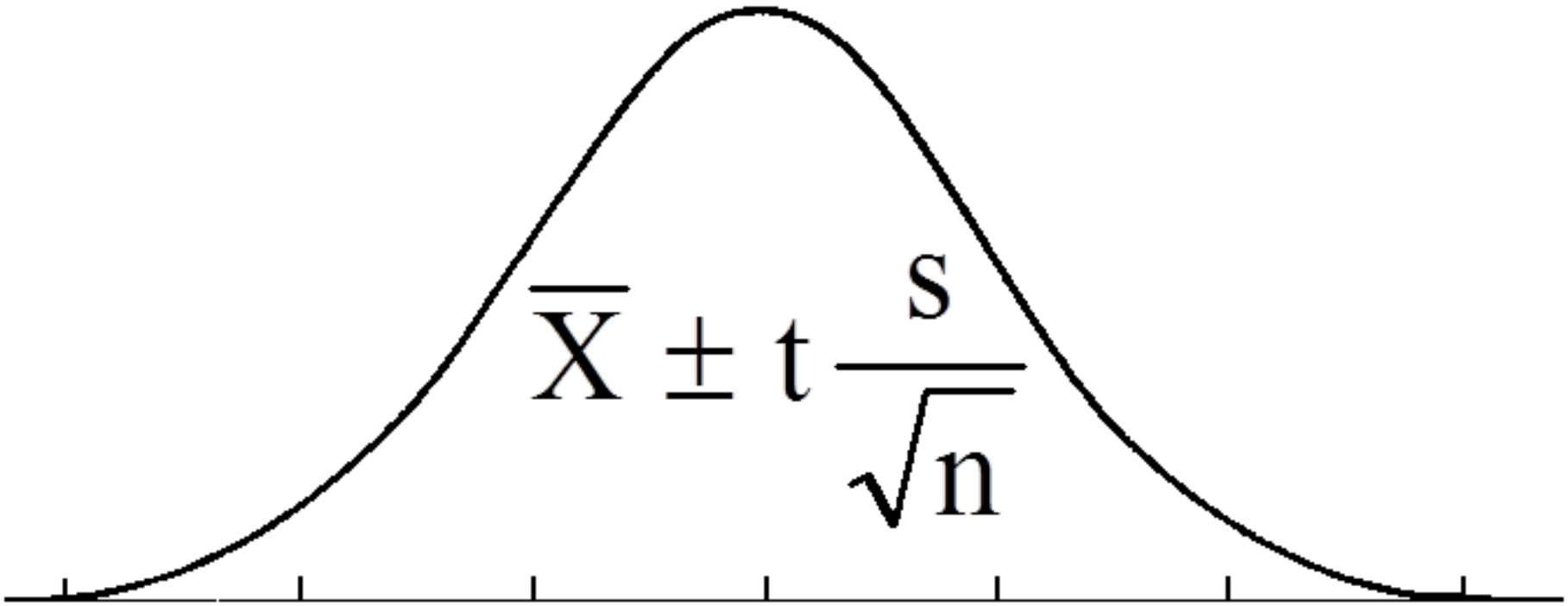
NORMAL DISTRIBUTION

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

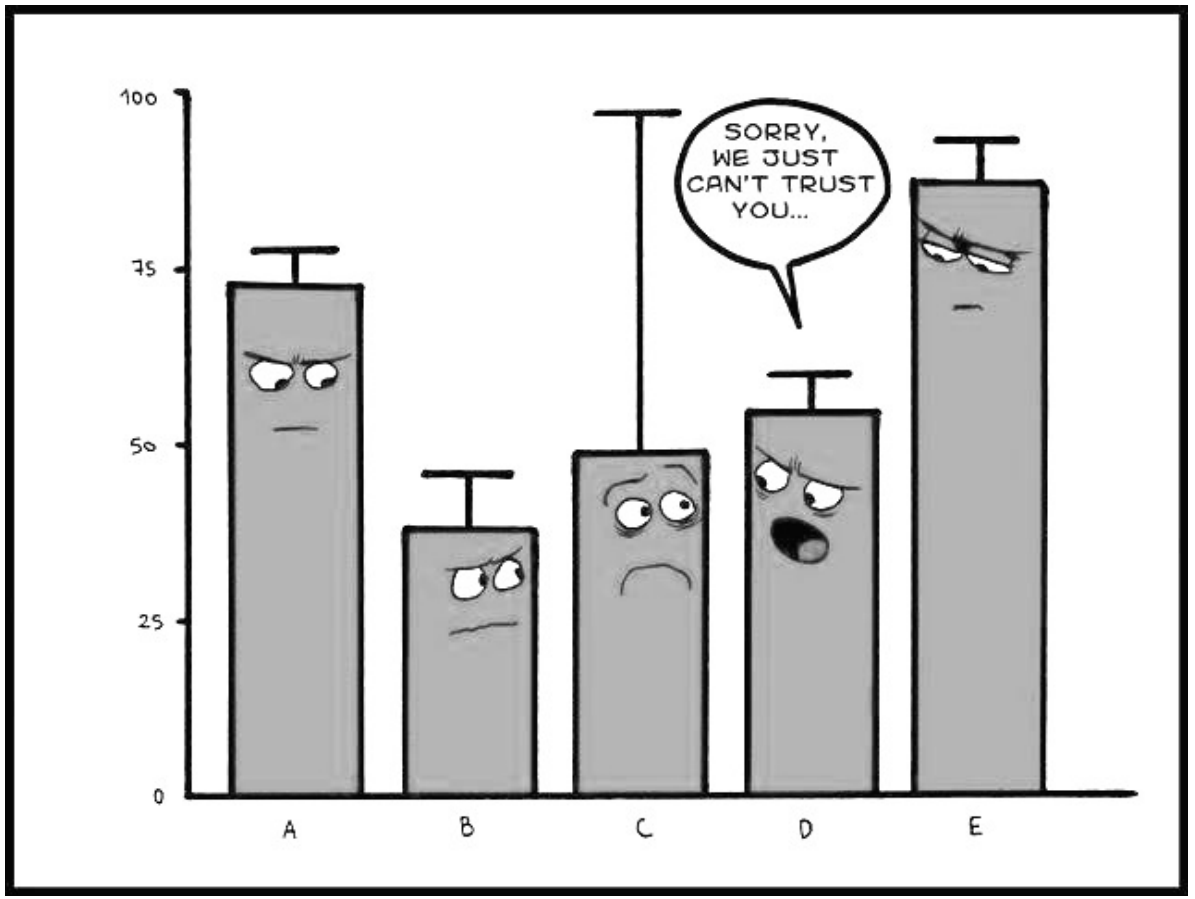
NORMAL DISTRIBUTION

"Exact" Confidence Intervals



t ~ 1.96 for large samples

CONFIDENCE INTERVALS



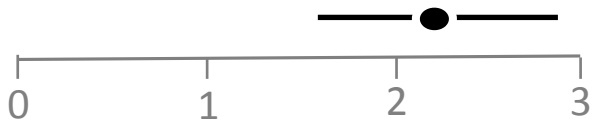
CONFIDENCE INTERVALS

- 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]

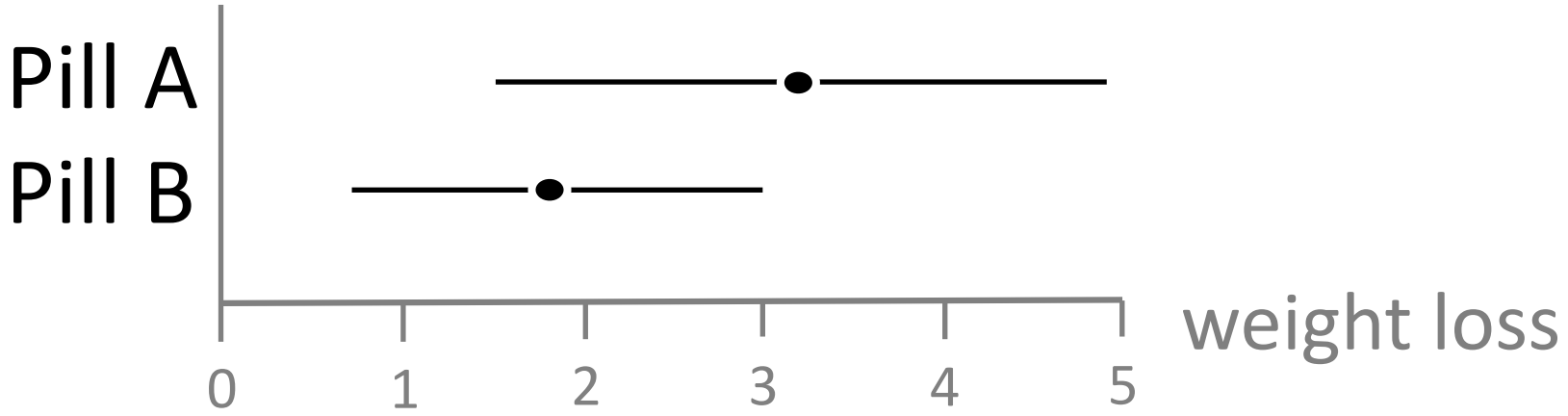
  2.2m +/- 0.6m

  from 1.6m to 2.8m



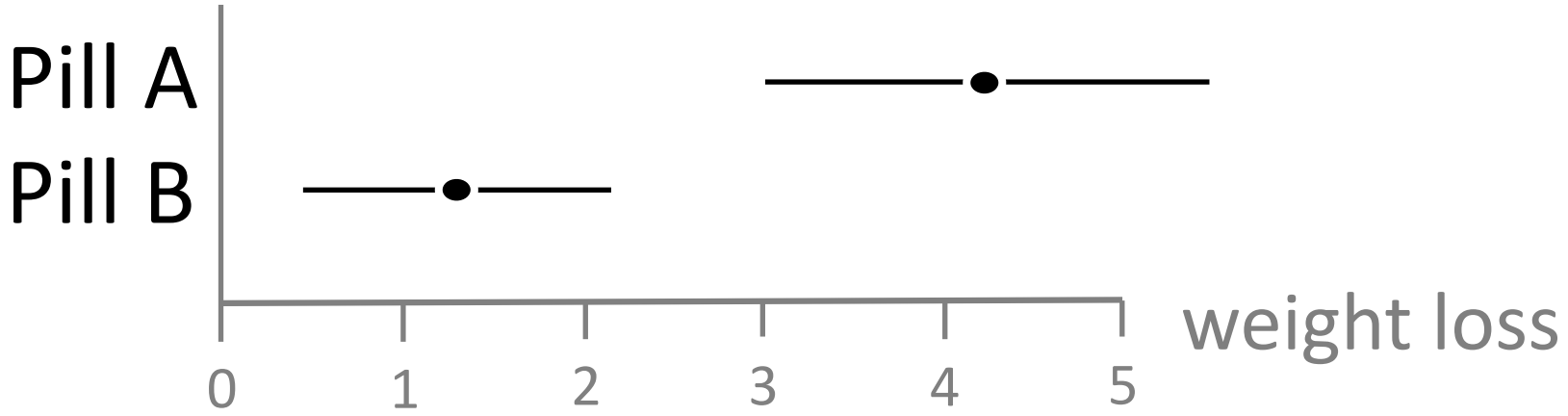
CONFIDENCE INTERVALS

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



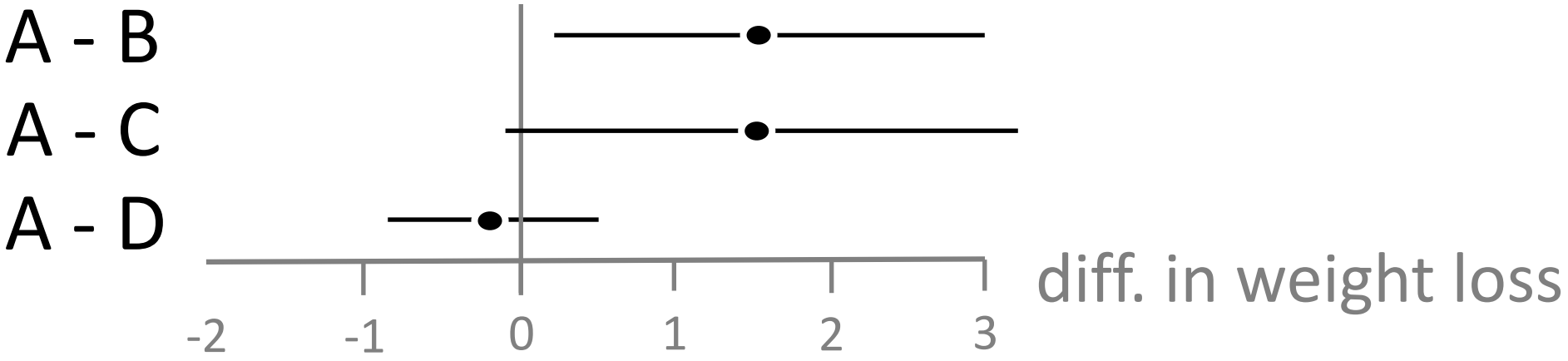
CONFIDENCE INTERVALS

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



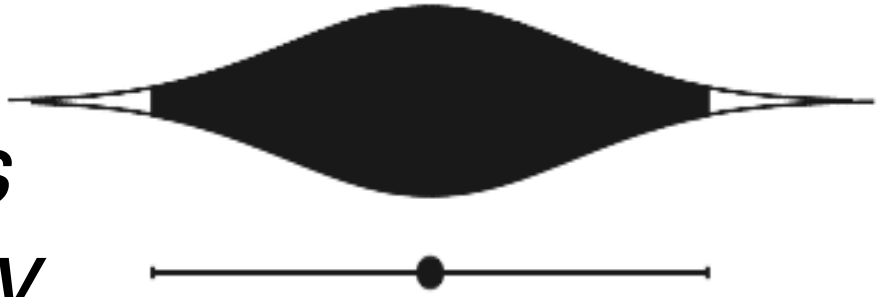
CONFIDENCE INTERVALS

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



CONFIDENCE INTERVALS

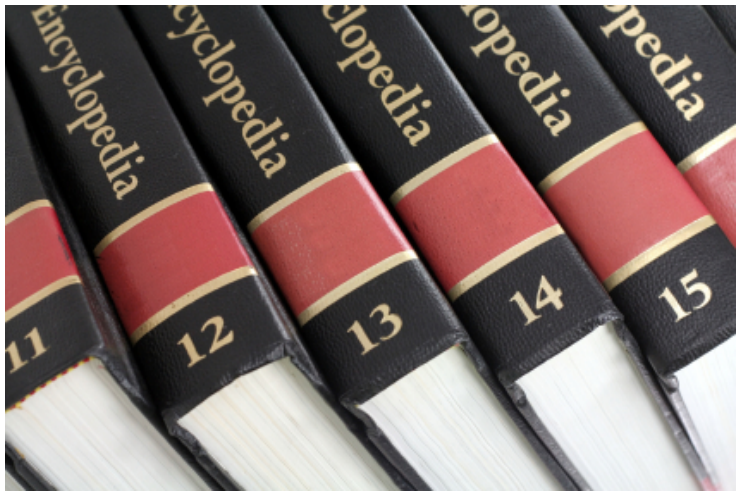
- "*values close to our M are the best bet for μ, and values closer to the limits of our CI are successively less good bets.*"



(Cumming, 2013)

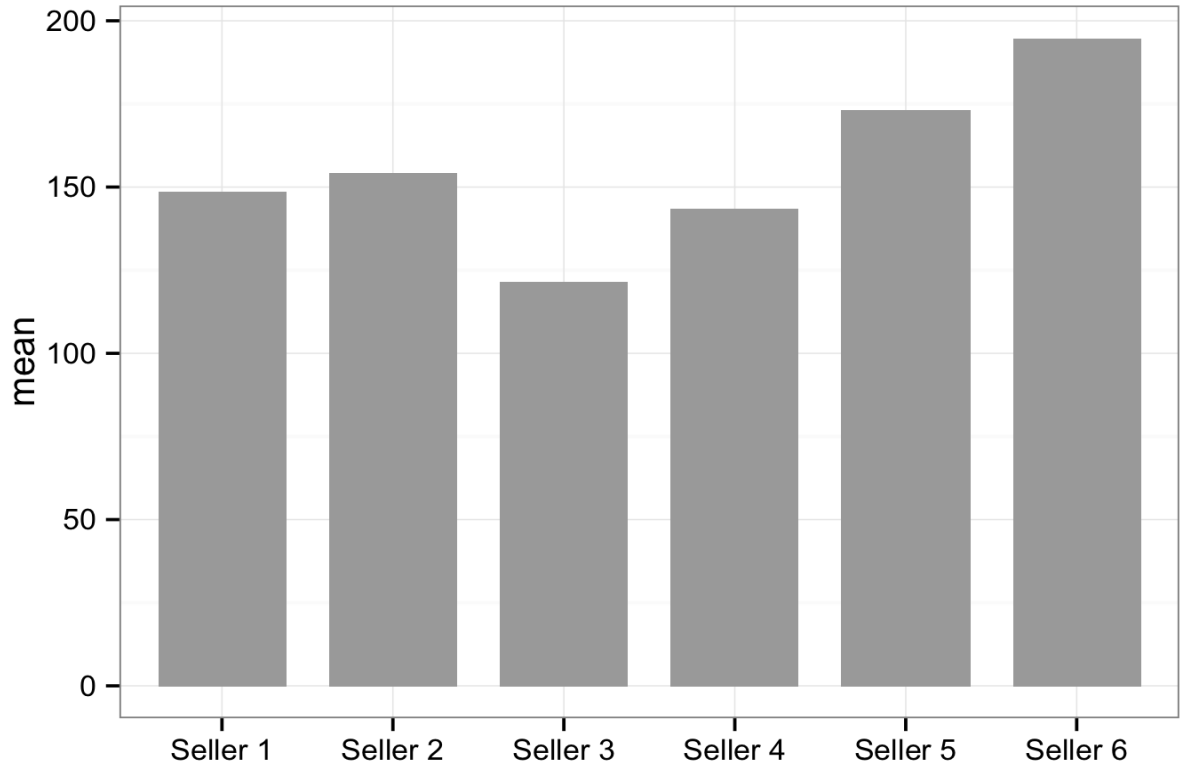
BACK TO OUR EXAMPLE

- Selling encyclopedias



Average Sales

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



https://www.lri.fr/~dragice/stats-rjc.zip

Doing the New Statistics

Pierre Dragicevic

7èmes Rencontres des Jeunes Chercheurs en IHM
Juin 2015

Statistics

- In the context of HCI user studies

- 3 things to keep in mind:
  - Not all papers need a user study
  - Not all user studies are experiments
  - Doing experiments is not all about statistics

- but doing experiments require doing some statistics

Statistics

We gave a data retrieval task to 12 subjects. Half of them used a bar chart and the other half used a line chart.

The measured accuracies were (12.1%, 11.6%, 18.3%, 19.2%, 11.1%, 7.0%) for bar charts, and (13.0%, 13.9%, 12.1%, 13.5%, 21.9%, 12.4%) for line charts.

Statistics

We gave a data retrieval task to 12 subjects. Half of them used a bar chart and the other half used a line chart.

Average accuracy was **9.2%** for bar charts, and **13.2%** for line charts.

Bad HCI Statistics

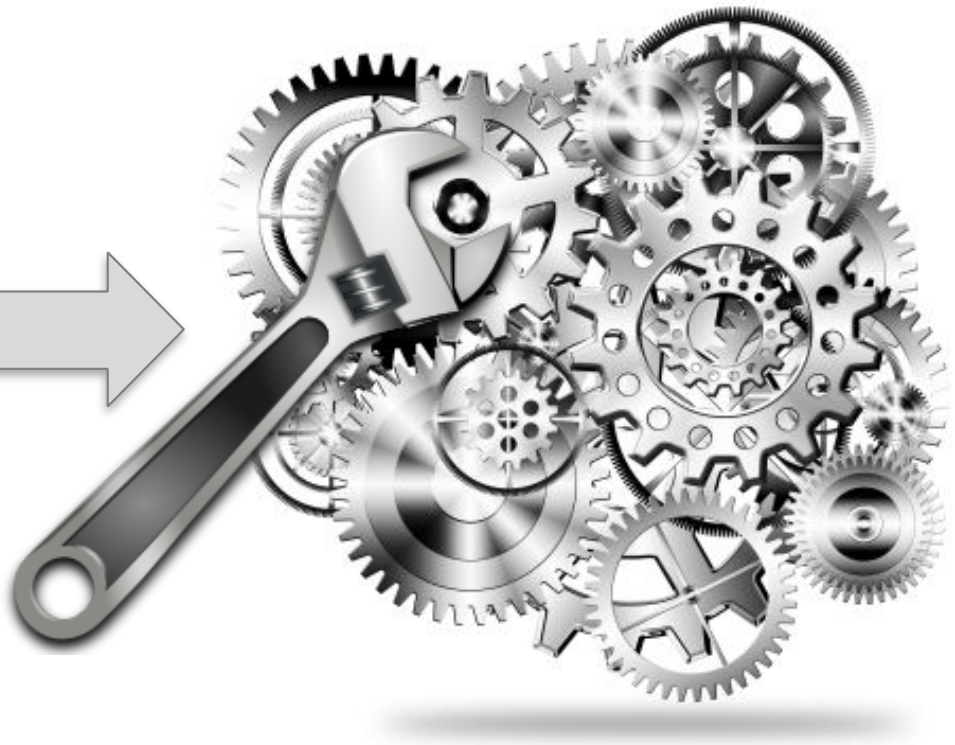
Experiment data

Technique works

Technique does not work

Bad HCI Statistics

Experiment data

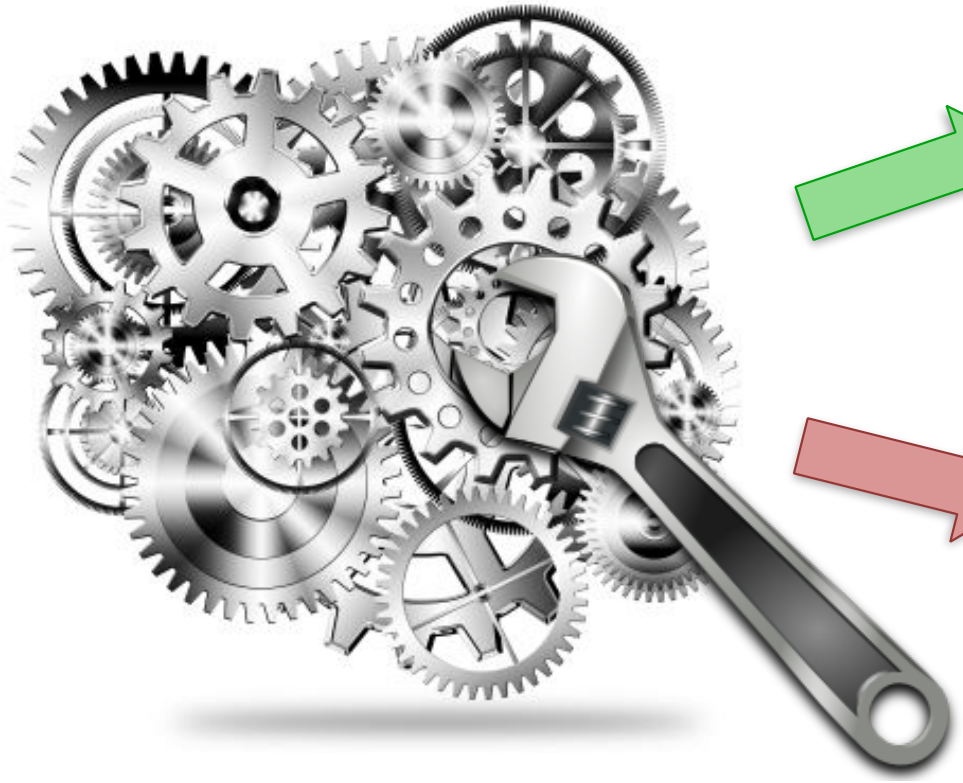


Technique works

Technique does not work

Bad HCI Statistics

Experiment data

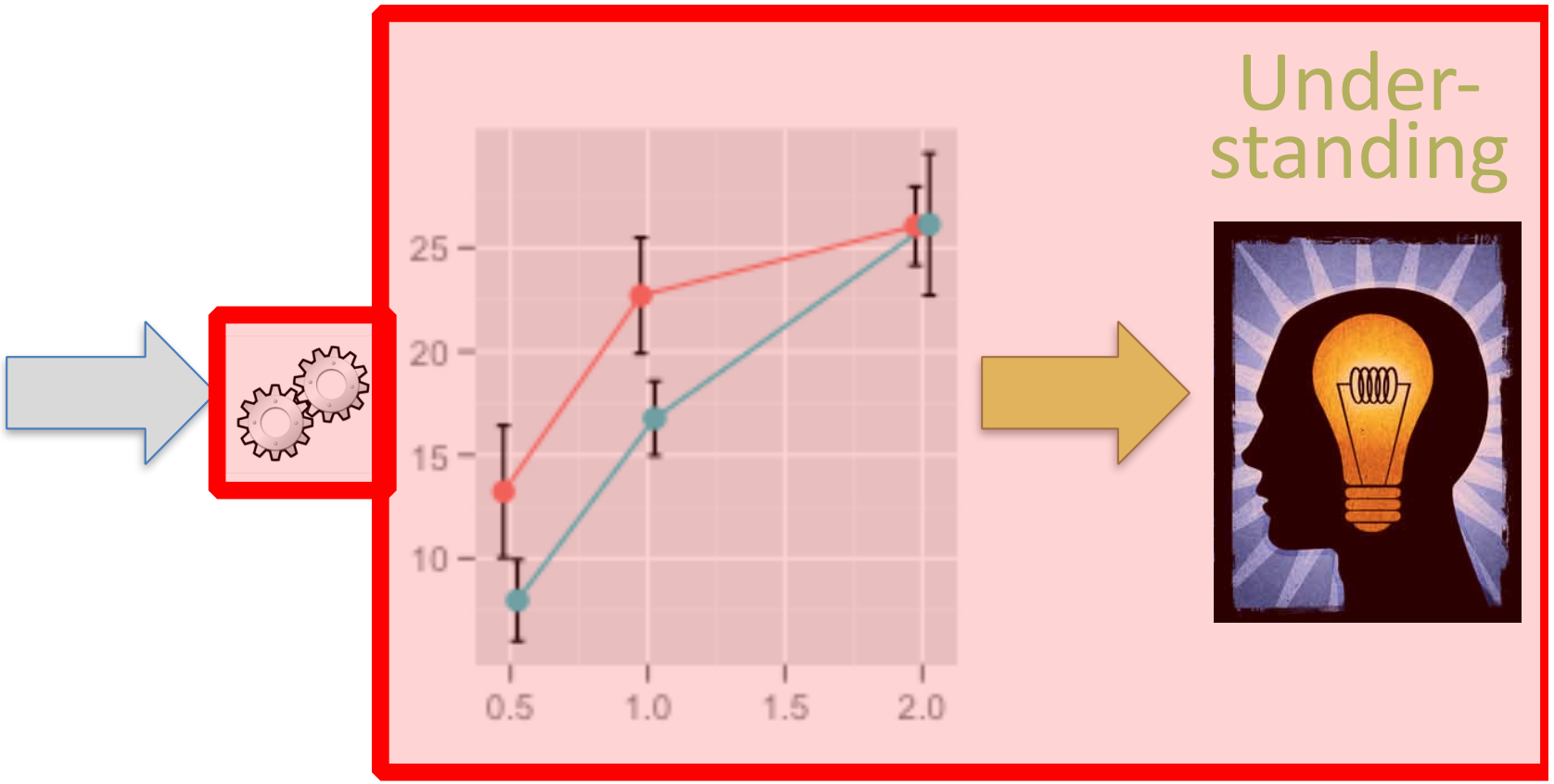


Technique works

Technique does not work

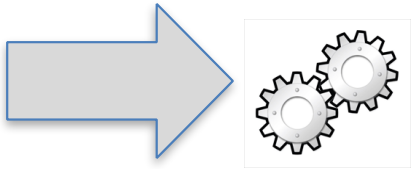
What HCI Statistics are for

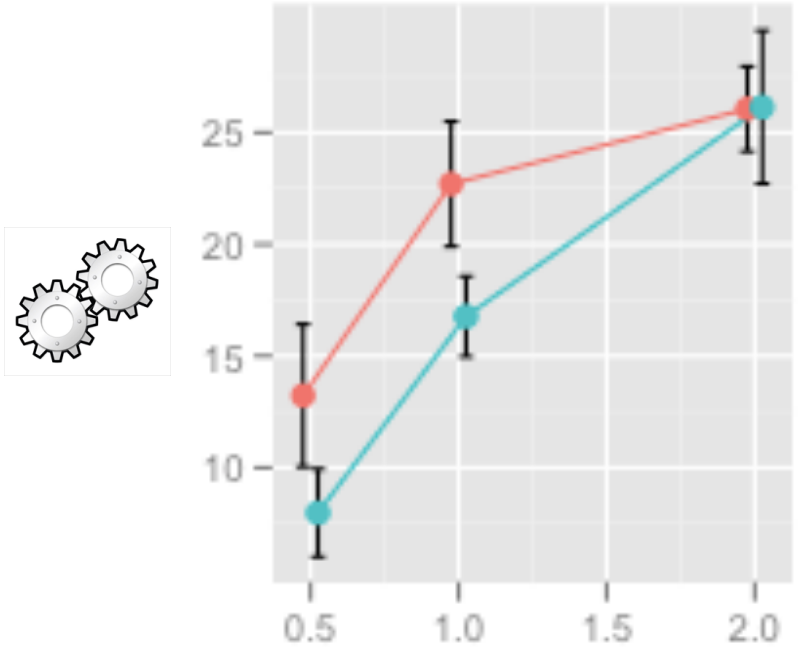
Experiment data



What HCI Statistics are for

Experiment data





Under-standing

What HCI Statistics are for

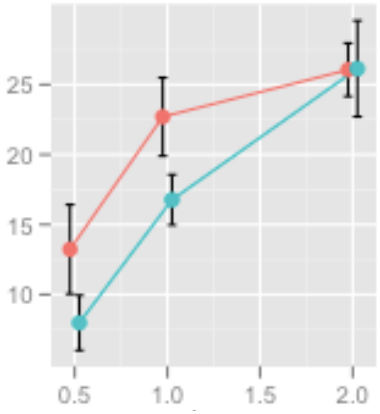
Under-
standing

Investigator

What HCI Statistics are for

Under-
standing

Investigator

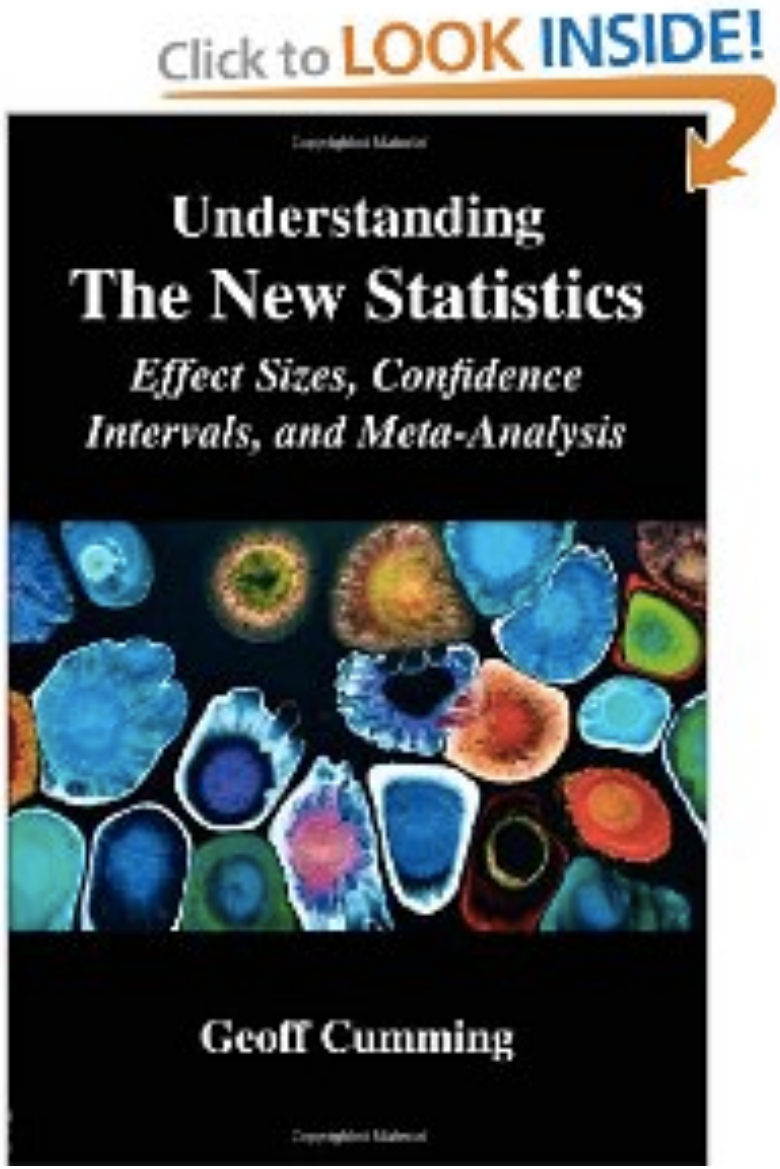


Publication

Understanding

Peers

The New Statistics



The New Statistics

« *The techniques **are not new**, but adopting them widely would be new for many researchers, as well as highly beneficial.* »

(Cumming, 2013)

The New Statistics

- The « old » statistics:
  - Null hypothesis significance testing (NHST)
  - *p*-values
  - Example: tech A is faster than tech B, $p = .032$
- The « new » statistics:
  - Estimations instead of tests
  - Effect sizes + confidence intervals (CIs)
  - Example: tech A is faster than tech B by 1.3 seconds, 95% CI [0.3s, 1.6s].

FAQ

- What's an effect size?
- What's a CI?
- Why switch to CIs? Who says that?
- Is reporting *p*-values + CIs OK?
- How to compute CIs?
- How to graph CIs?
- How to interpret CIs?
- Will my paper be rejected?

References

- **(Keene, 1995)** The log transform is special
- **(Schmidt and Hunter, 1997)** Eight common but false objections to the discontinuation of significance testing in the analysis of research data.
- **(Wilkinson et al, 1999)** Statistical Methods in Psychology Journals.
- **(Cumming and Finch, 2005)** Inference by Eye: Confidence Intervals and How to Read Pictures of Data.
- **(Baguley, 2009)**. Standardized or simple effect size: What should be reported?
- **(Sauro and Lewis, 2010)**. Average task times in usability tests: what to report?
- **(Cumming, 2011)**. Cumming, G. Understanding the New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis.
- **(Dragicevic, 2012)**. My Technique is 20% Faster: Problems with Reports of Speed Improvements in HCI.
- **(Kirby and Gerlanc, 2012)**. BootES: An R Package for Bootstrap Confidence Intervals on Effect Sizes
- **(Cumming, 2013)** The New Statistics: Why and How.

What's an effect size?

- Taken broadly, « *the amount of something that might be of interest* » (Cumming, 2011)
- E.g., writing « tech A is faster than tech B by 1.3 seconds » is reporting an effect size
- Things like Cohen's *d* are *standardized* effect sizes
- Many recommend reporting simple (unstandardized) effect sizes

What's an effect size?

" Only rarely will uncorrected standardized effect size be more useful than simple effect size. It is usually ***far better to report simple effect size*** [...] *"*

(Baguley, 2009)

What's an effect size?

*" If the **units of measurement are meaningful** on a practical level (e.g., number of cigarettes smoked per day), then we usually prefer an **unstandardized measure** (regression coefficient or mean difference) to a standardized measure (r or d). "*

(Wilkinson et al., 1999)

What's an effect size?

*"(i) a **preference for simple effect size** over standardized effect size, and*

*(ii) the **use of confidence intervals** to indicate a plausible range of values the effect might take."*

(Baguley, 2009)

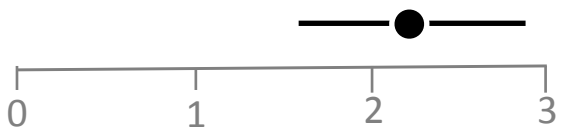
What's a confidence interval?

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

  2.2 sec, 95% CI [1.6, 2.8]

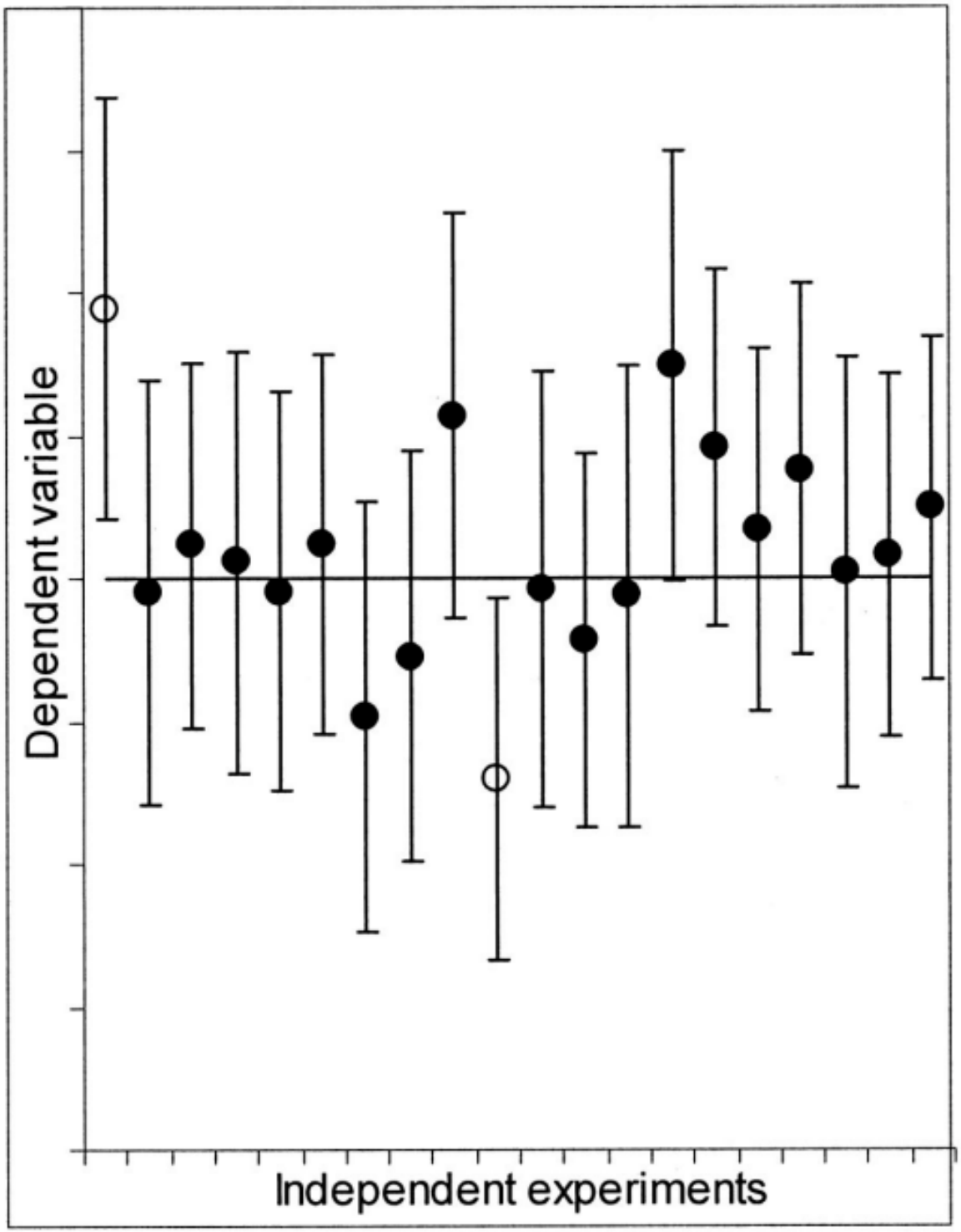
  2.2 sec +/- 0.6

  from 1.6 to 2.8 sec



What’s a confidence interval?

- Cumming’s favorite interpretation
- « *our CI is just one from an infinite sequence* »
  (Cumming and Finch, 2005)



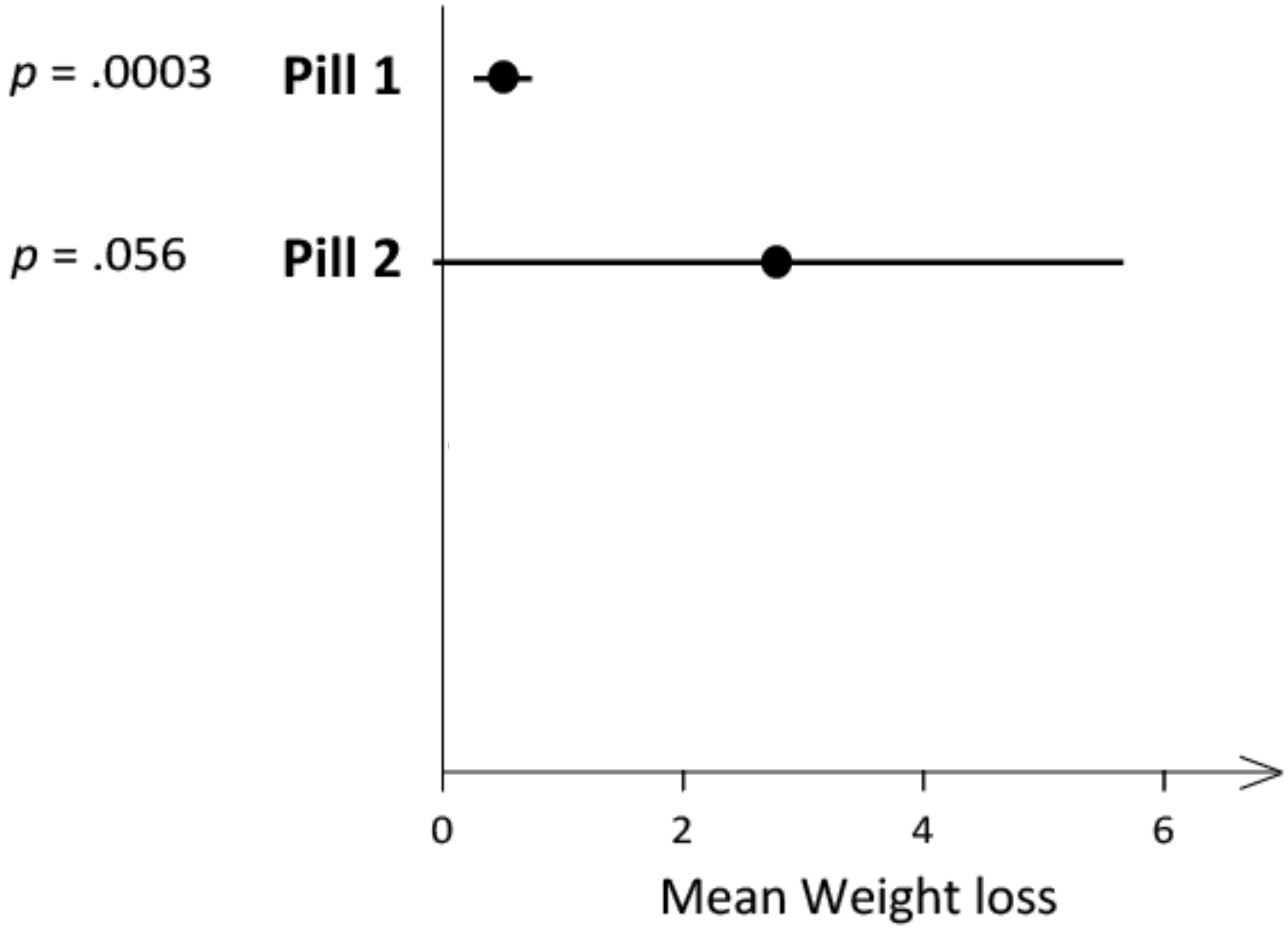
- Make sure you check *the dance of p-values* on youtube

Why Switch to CIs? Who says that?

- 300+ articles by renown methodologists have been questioning NHST since the 1950s
- Many recommend switching to estimation
- Researchers have been mostly ignoring them, but now things seem to be changing
- The problem with NHST is mostly a *human factor* problem, so we should know more!

- More at www.aviz.fr/badstats

Which weight-loss pill would you recommend?

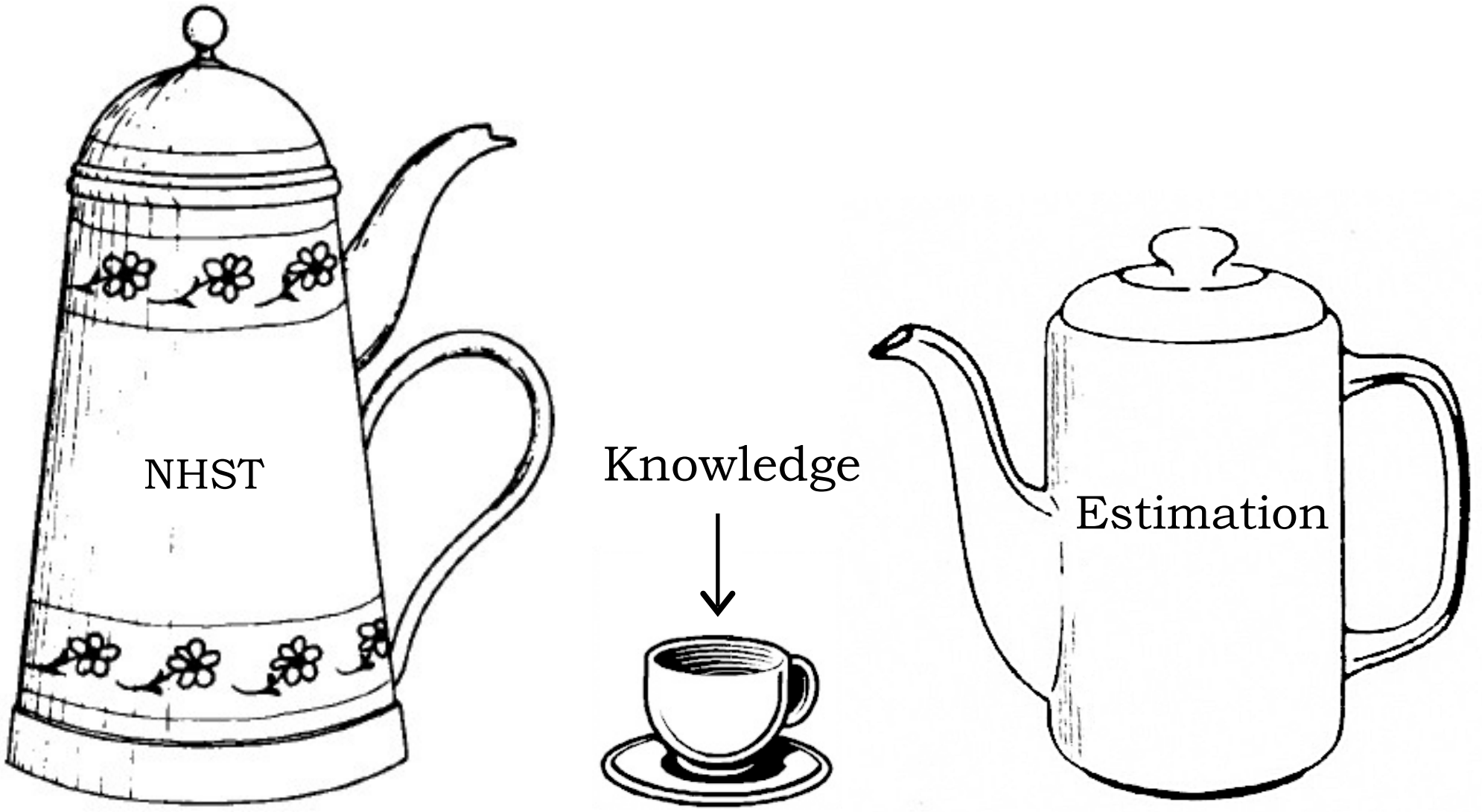


Error bars are 95% CIs

p-values are based on a null hypothesis of no effect

Is reporting *p*-values + CIs OK?

Is reporting *p*-values + CIs OK?



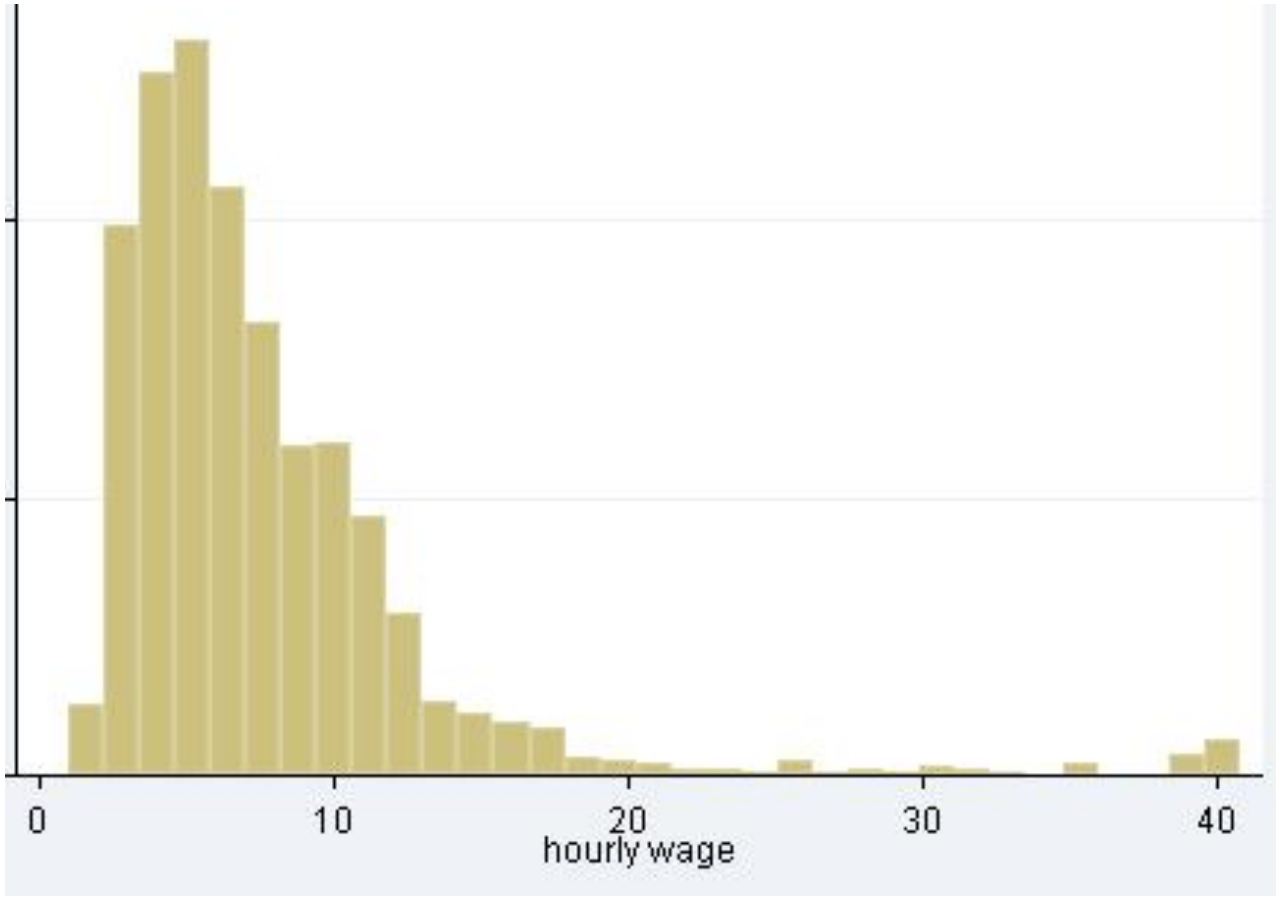
Is reporting *p*-values + CIs OK?

How to Compute CIs?

- Relatively easy using the R package
- A good place to start: tinyurl.com/r-ci-tutorial
- Aggregate your data first!!
- Less resources on:
  - Non-normal distributions
  - Complex designs
    - i.e., anything else than a between-subjects design with one factor and two levels

Non-Normal Distributions

- Skewed distributions



Non-Normal Distributions

“ For continuous positive data measured on an interval scale, a ***log transformed analysis should frequently be preferred*** *to an untransformed analysis. No special justification beyond that sufficient to support an untransformed analysis should be required from the data obtained. ”*

(Keene, 1995)

Non-Normal Distributions

- Log transformation
  - Transform all your raw time measurements into logs
  - Do all your stats
  - Transform back when presenting your results

- Two important things:
  - Arithmetic means become geometric means
  - Differences between means become ratios between geometric means

Non-Normal Distributions

- Logarithmic identities

$$\log_b(xy) = \log_b(x) + \log_b(y)$$

$$\log_b(x^d) = d \log_b(x)$$

Non-Normal Distributions

- *data* = (a, b, c)
- *logdata* = (log(a), log(b), log(c))
- *mean(logdata)* = (log(a) + log(b) + log(c)) / 3
- *antilog(mean(logdata))*
  = exp [(log(a) + log(b) + log(c)) / 3]
  = exp [log(abc) / 3]
  = exp [(1/3) * log(abc)]
  = exp [log((abc)$^{1/3}$)]
  = abc$^{1/3}$

Non-Normal Distributions

- *antilog(mean(logdata))*
  = n^{th} root of the product of all measurements

Non-Normal Distributions

- *antilog(mean(logdata))*
  = n^{th} root of the product of all measurements

- *" The **geometric mean** is defined as the n^{th} root of the product of n numbers."* Wikipedia

$$\left(\prod_{i=1}^{n} a_i\right)^{1/n} = \sqrt[n]{a_1 a_2 \cdots a_n}.$$

Non-Normal Distributions

“ When providing an estimate of the average task time for small sample studies (n<25), the ***geometric mean*** *is the best estimate of the center of the population (the median).”*

(Sauro and Lewis, 2010)

Non-Normal Distributions

*“ To find the **geometric mean**, convert the raw times using **a log-transformation**, find the mean of the transformed data, then transform back to the original scale by exponentiating.”*

(Sauro and Lewis, 2010)

Non-Normal Distributions

- Log transformation
  - Transform all your raw time measurements into logs
  - Do all your stats
  - Transform back when presenting your results

- Two important things:
  - Arithmetic means become geometric means
  - Differences between means become ratios between geometric means

Non-Normal Distributions

- Other distributions
  - Exponential
  - With both lower and upper bound

- Use bootstrapping (resampling)
  - R package *boot*
  - Simple and works with about any distribution

  (Kirby and Gerlanc, 2012)

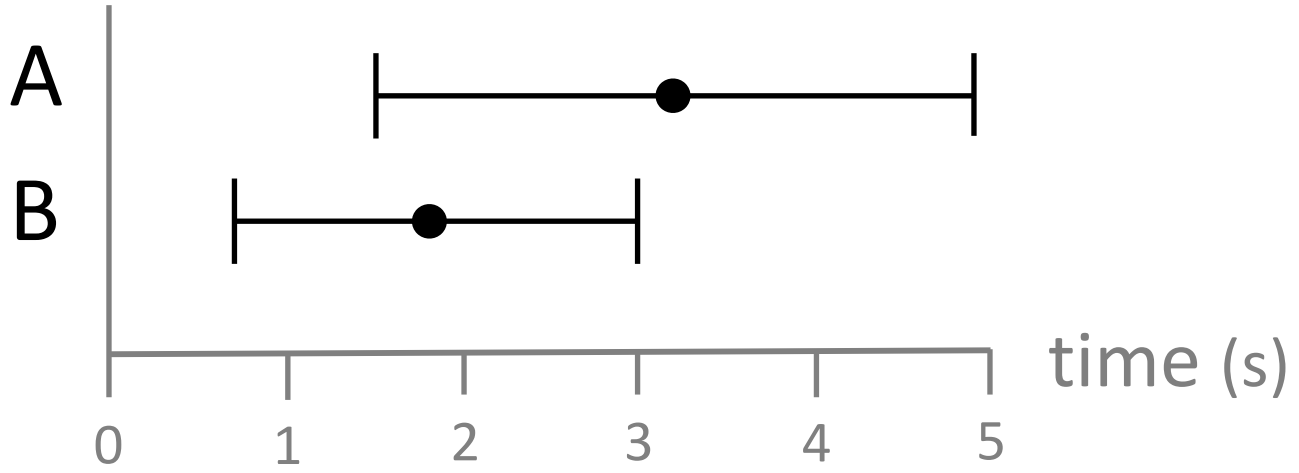
CIs on Differences

- Between-subject designs
  - different formula than sample mean
- Within-subject designs
  - just compute the difference on each pair of data points

- Multiple factors
- Multiple levels
- We'll see these later

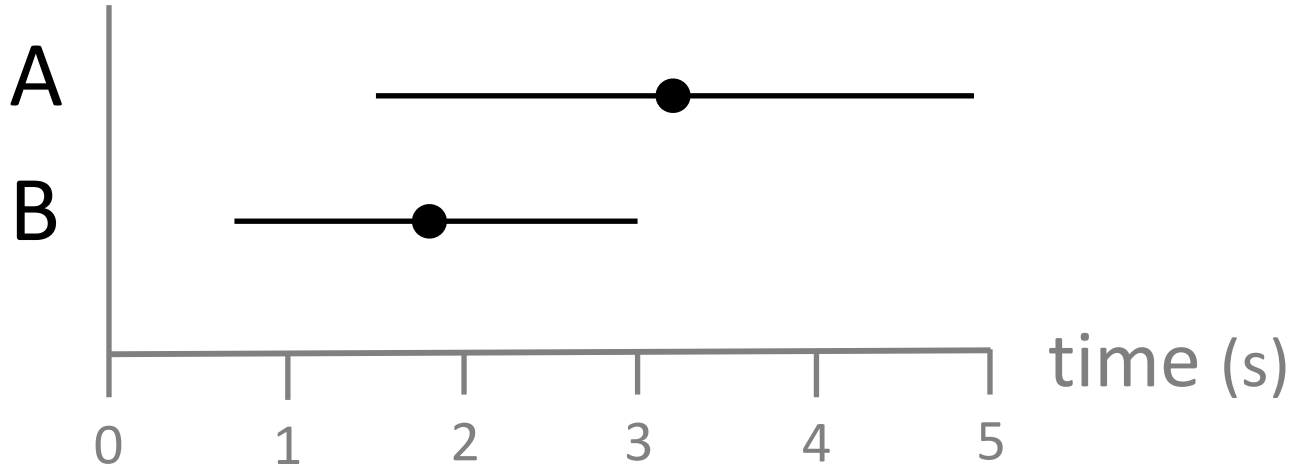
How to Graph CIs?

- As error bars



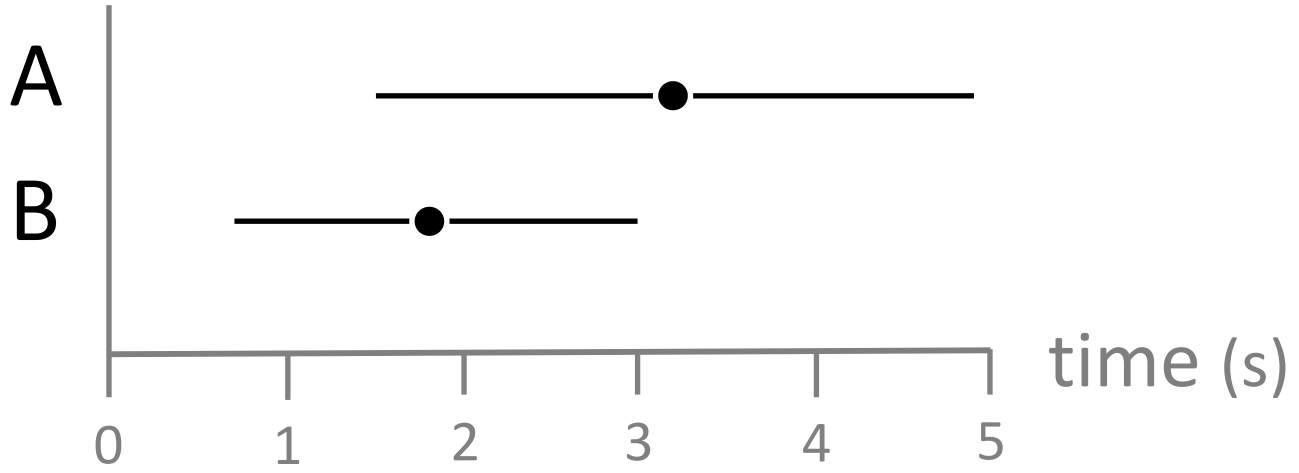
How to Graph CIs?

- As error bars
  - Better way:



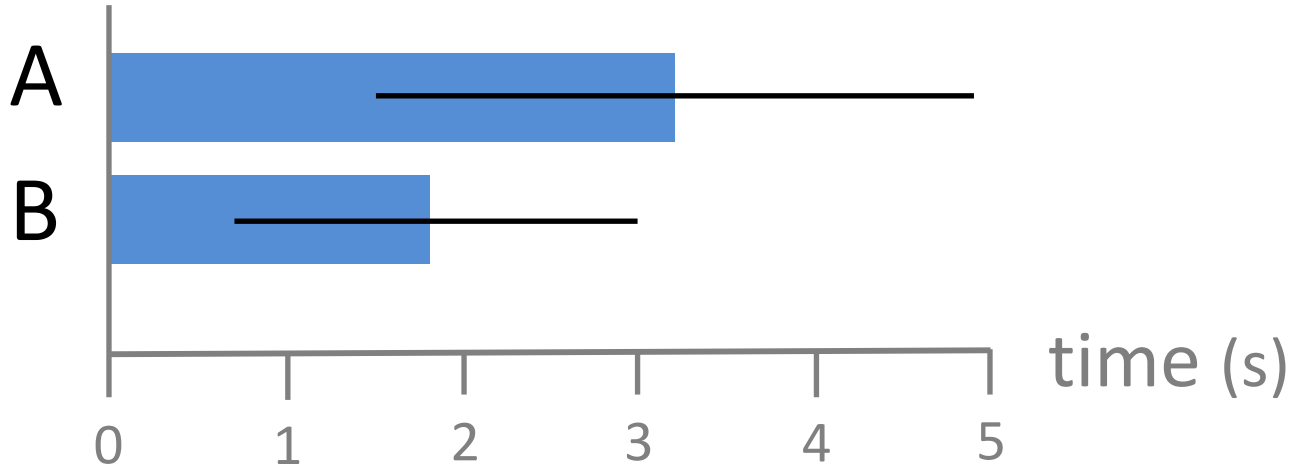
How to Graph CIs?

- As error bars
  - Slightly nicer:



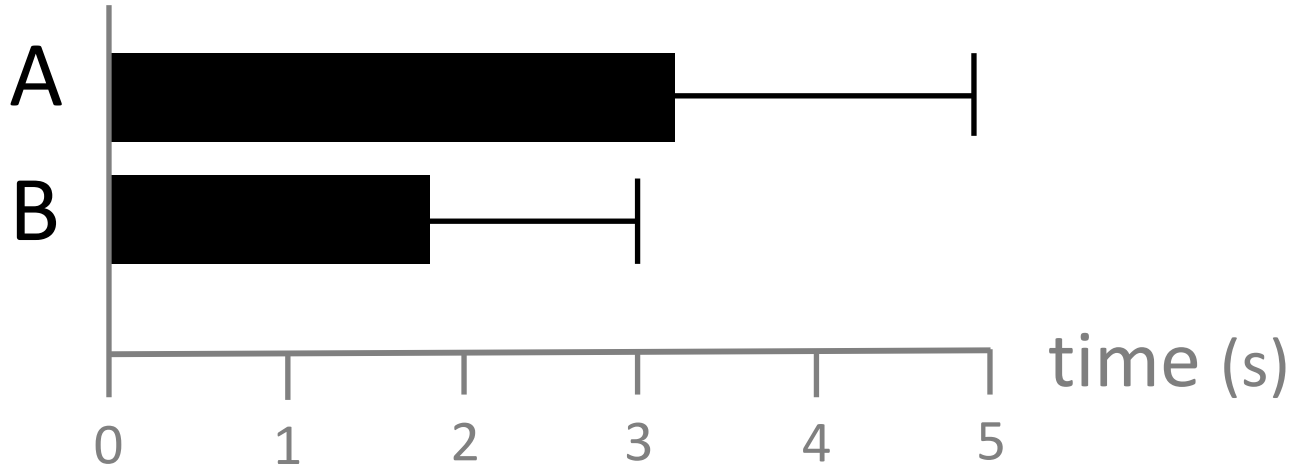
How to Graph CIs?

- As error bars
  - With bar charts:



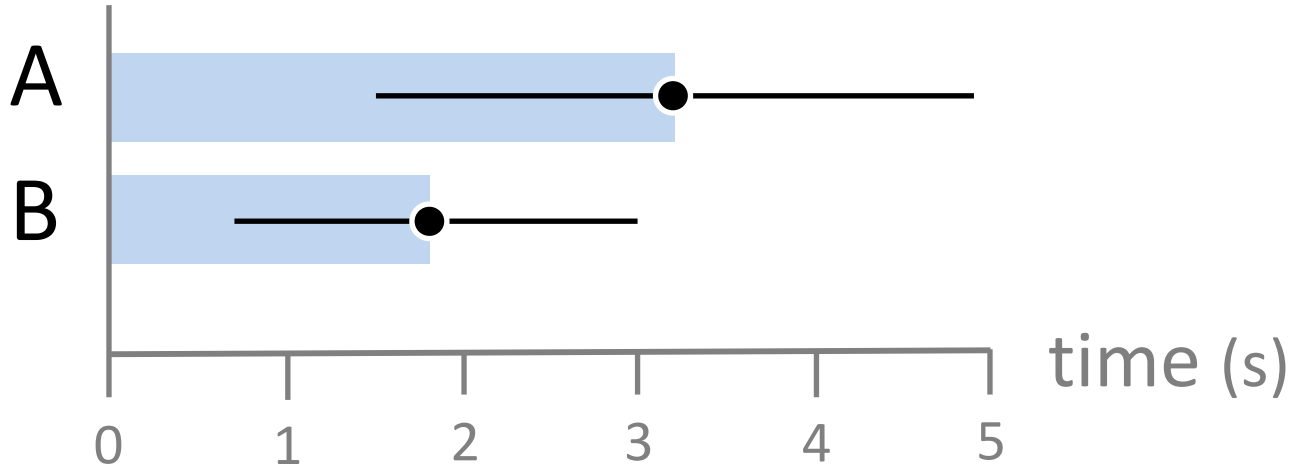
How to Graph CIs?

- As error bars
  - Dynamite plots:



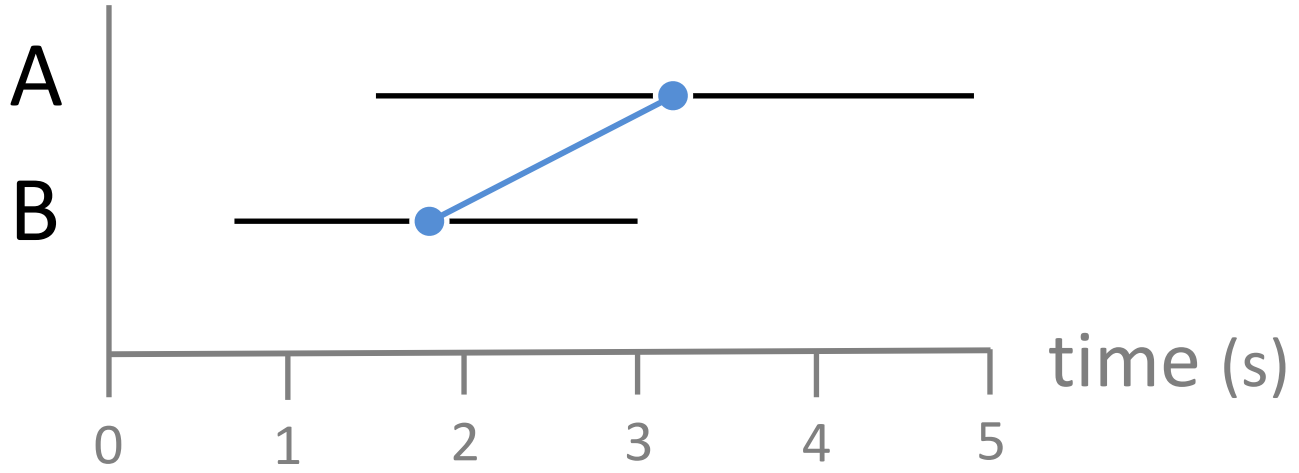
How to Graph CIs?

- As error bars
  - Perhaps a better approach:

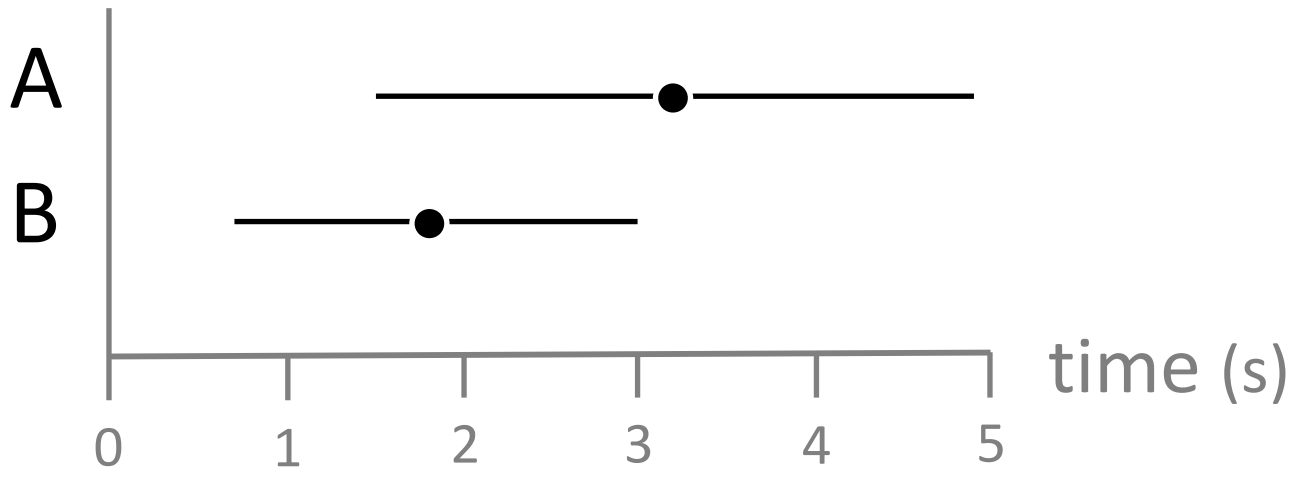


How to Graph CIs?

- As error bars
  - With line charts:

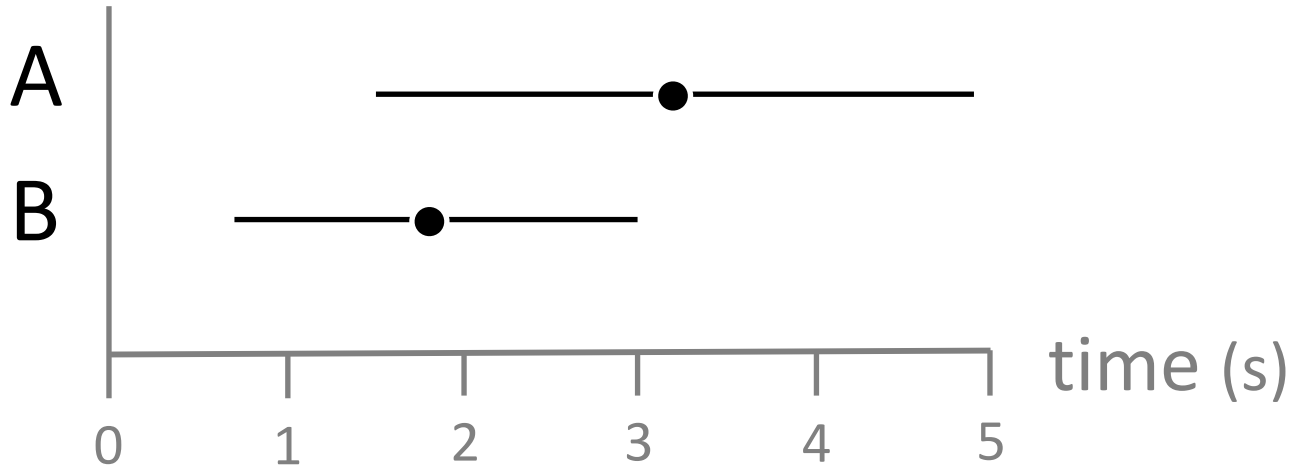


How to Interpret CIs?



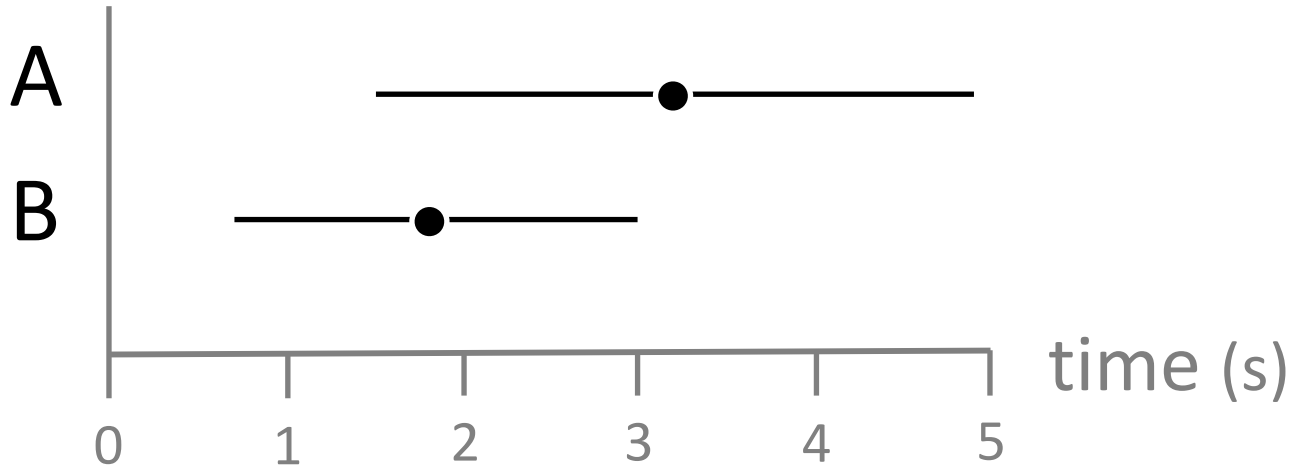
How to Interpret CIs?

- Error bars could be anything
  - Standard Error (SE), Variance, various CIs, etc.
  - Use 95% CIs and specify in the legend



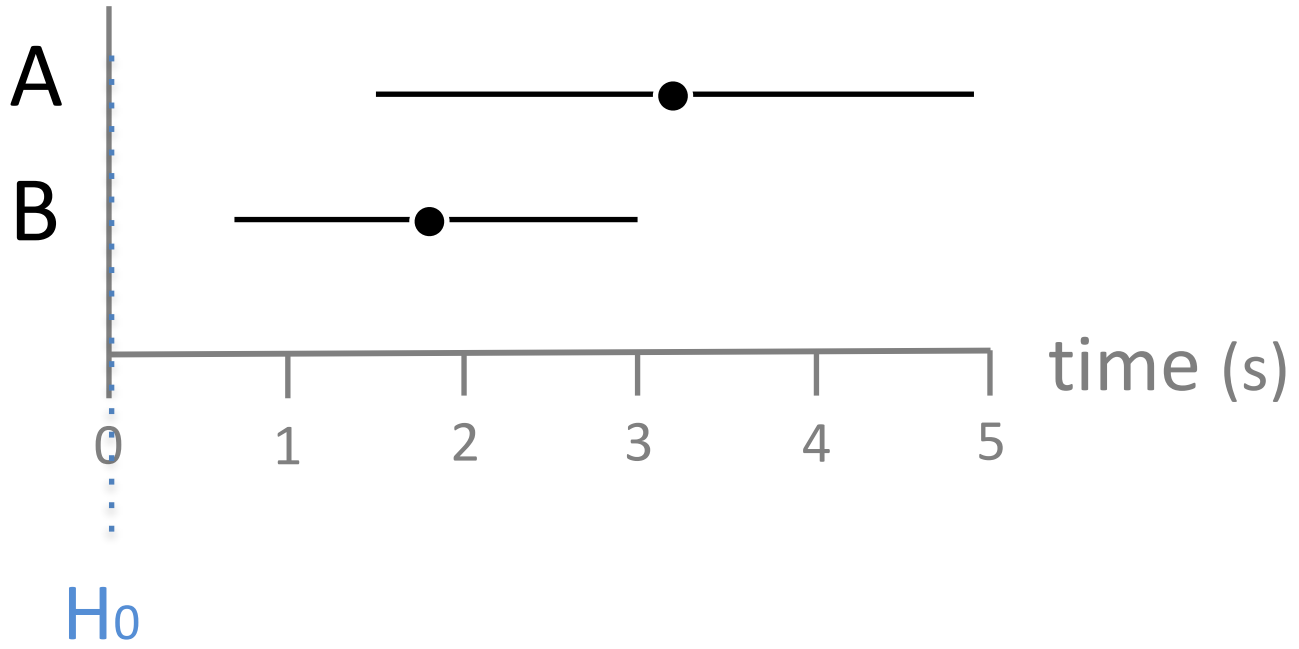
How to Interpret CIs?

- Null hypothesis: H_0 = 0s



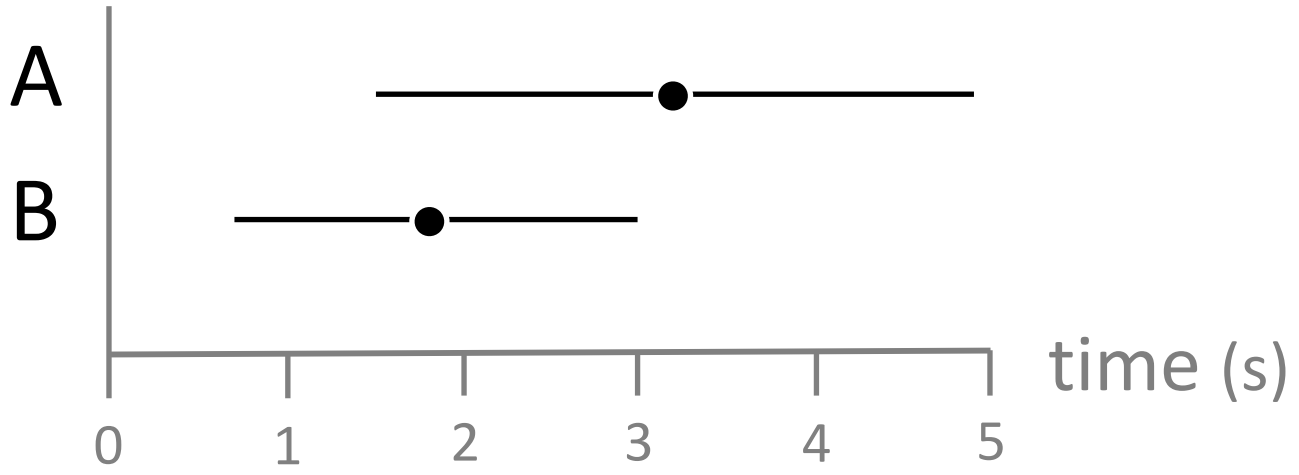
How to Interpret CIs?

- Null hypothesis: H_0 = 0s
  - For A, time is significantly different from 0s, $p < .05$
  - For B, time is significantly different from 0s, $p < .05$



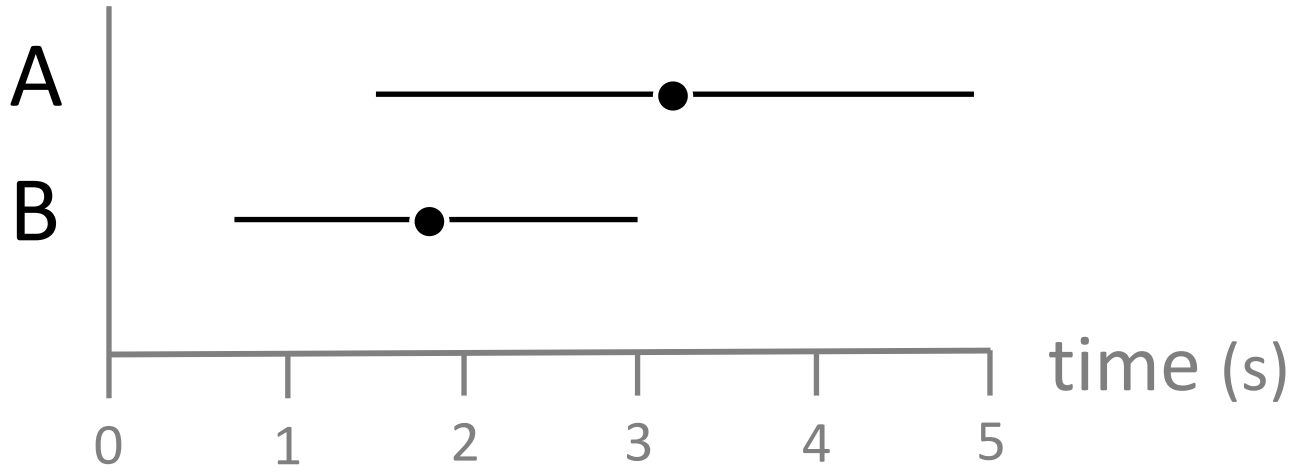
How to Interpret CIs?

- Null hypothesis: H_0 = 1s



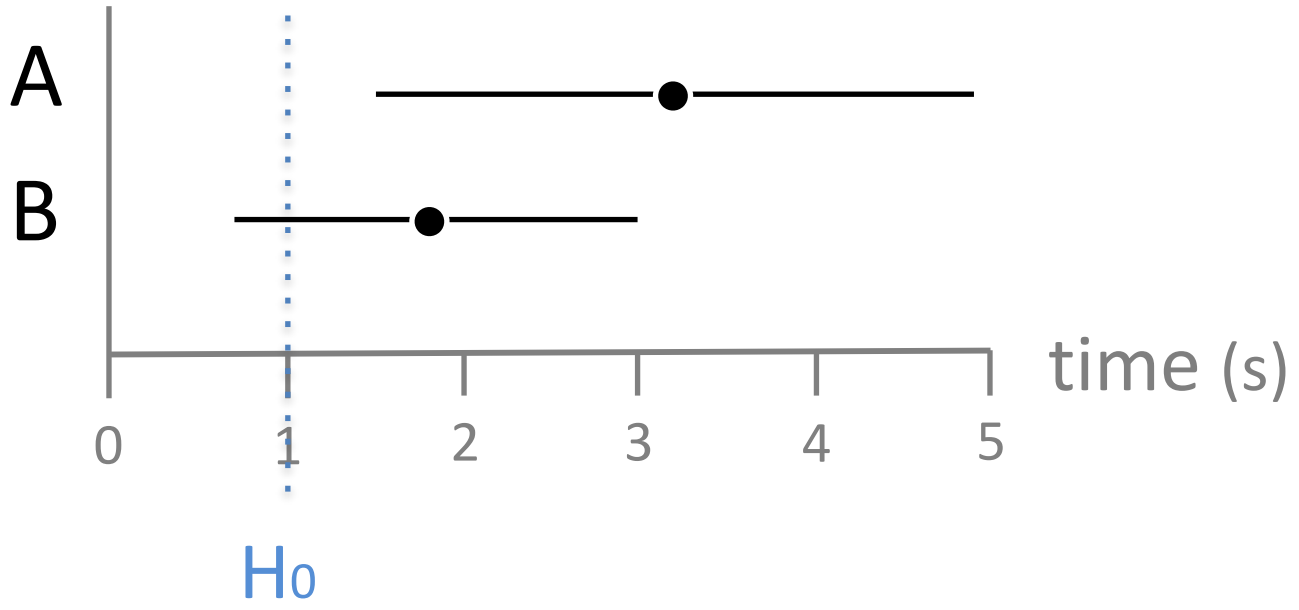
How to Interpret CIs?

- Null hypothesis: H_0 = 1s
  - For A, time is significantly different from 1s, $p < .05$
  - For B, time is not sig. different from 1s, ($p > .05$)



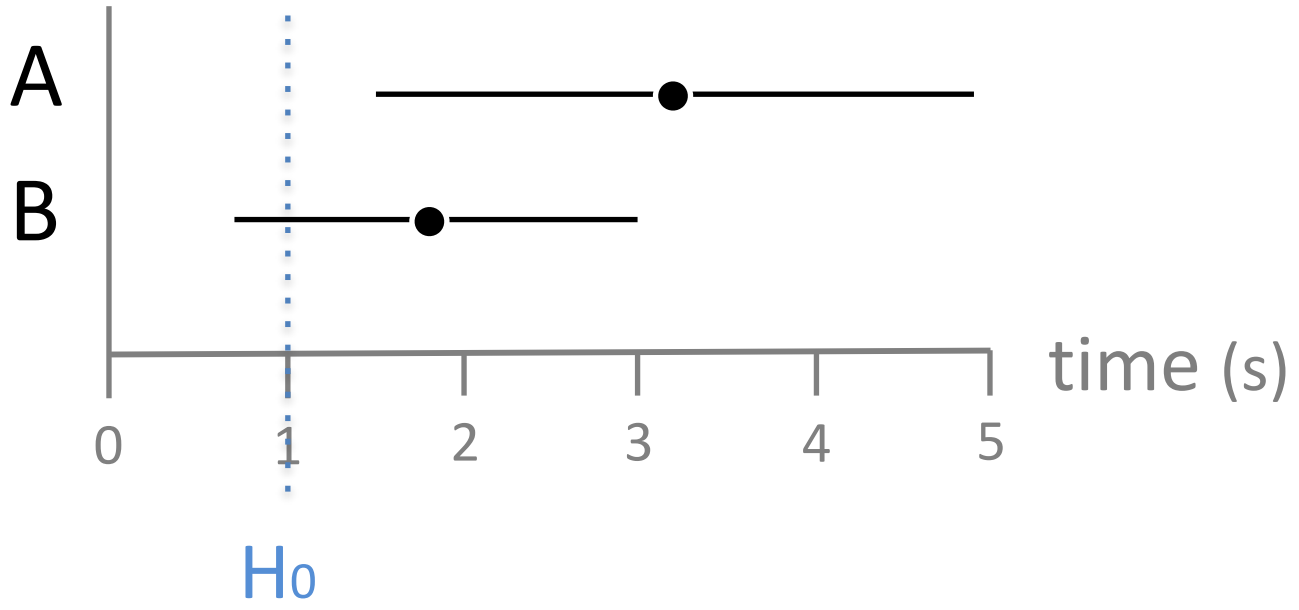
How to Interpret CIs?

- Null hypothesis: H_0 = 1s
  - For A, time is significantly different from 1s, $p < .05$
  - For B, time is not sig. different from 1s, $(p > .05)$



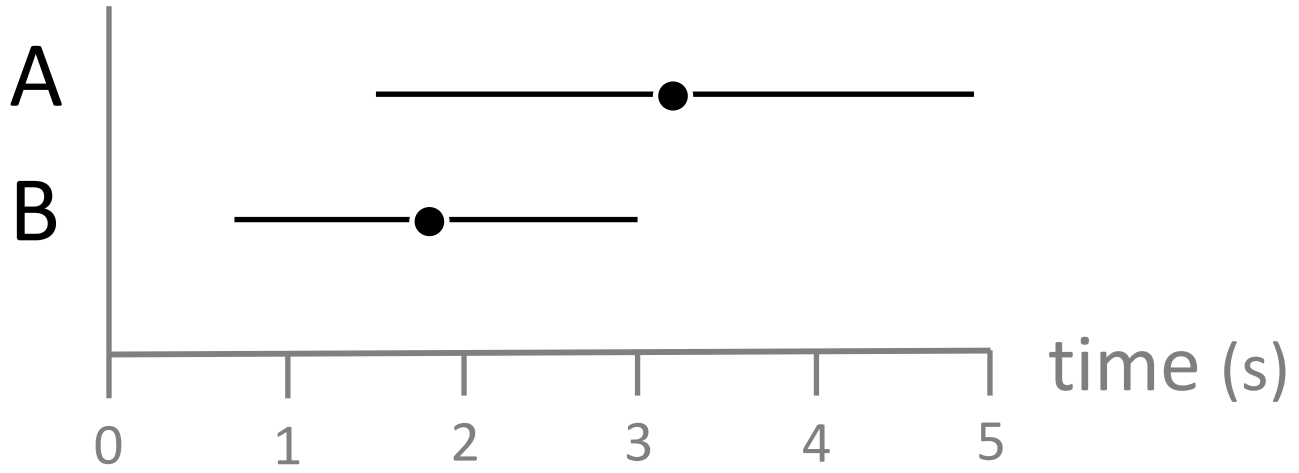
How to Interpret CIs?

- « *This is my least preferred way to interpret a CI: I earlier cited evidence that CIs can prompt better interpretation if NHST is avoided.* »
  (Cumming and Finch, 2005)



How to Interpret CIs?

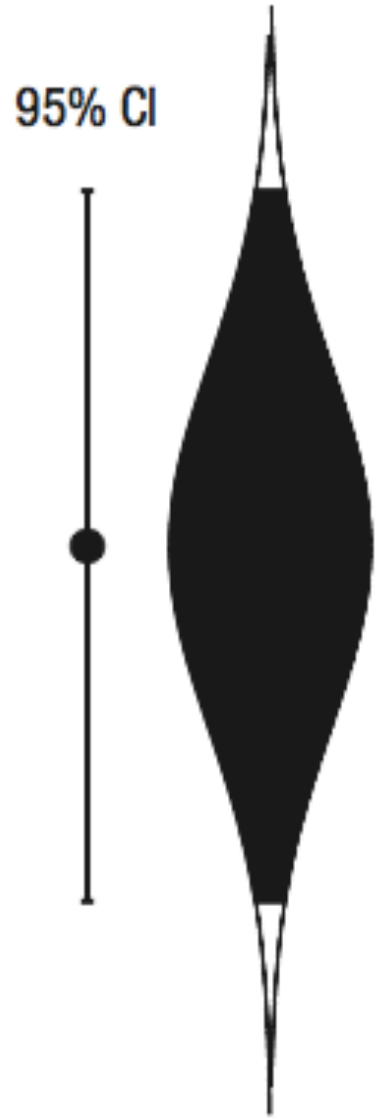
- « *a range of plausible values for μ. Values outside the CI are relatively implausible.* »
  (Cumming and Finch, 2005)



How to Interpret CIs?

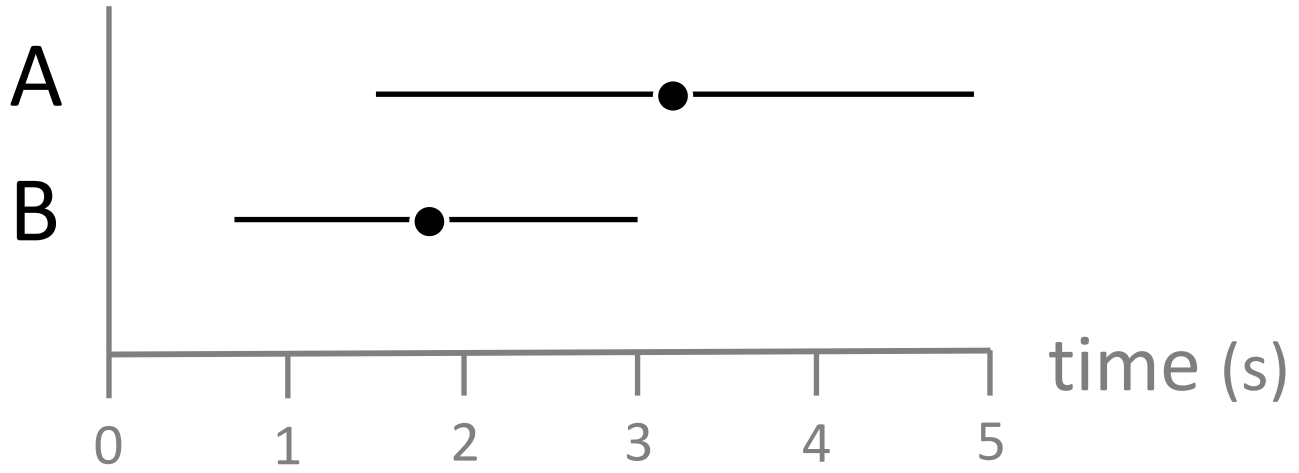
- *"values close to our M are the best bet for μ, and values closer to the limits of our CI are successively less good bets."*

(Cumming, 2013)



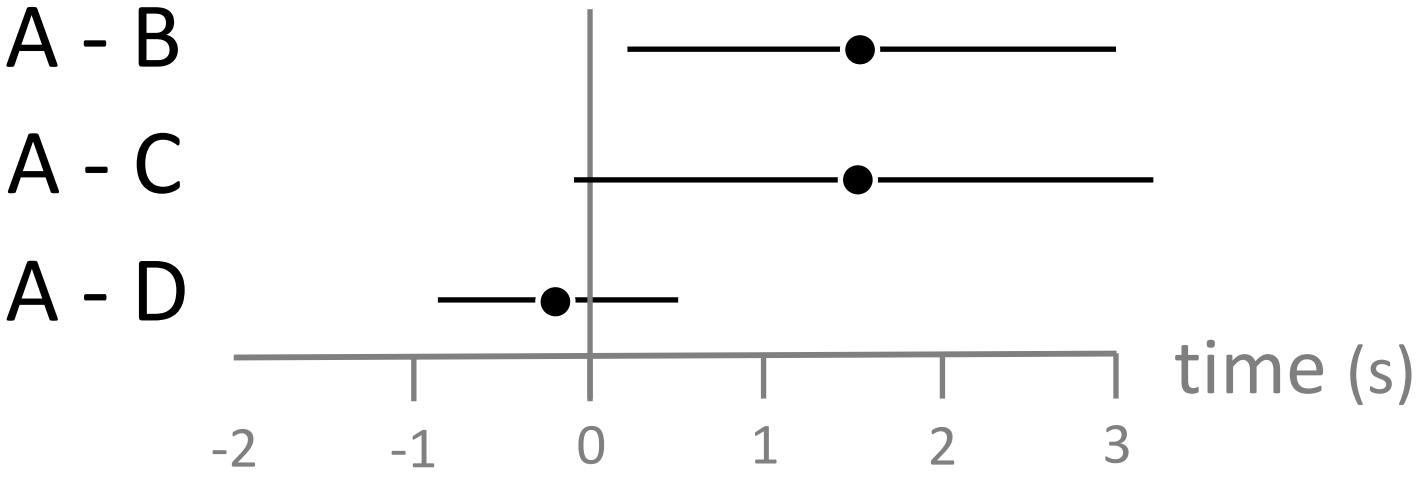
How to Interpret CIs?

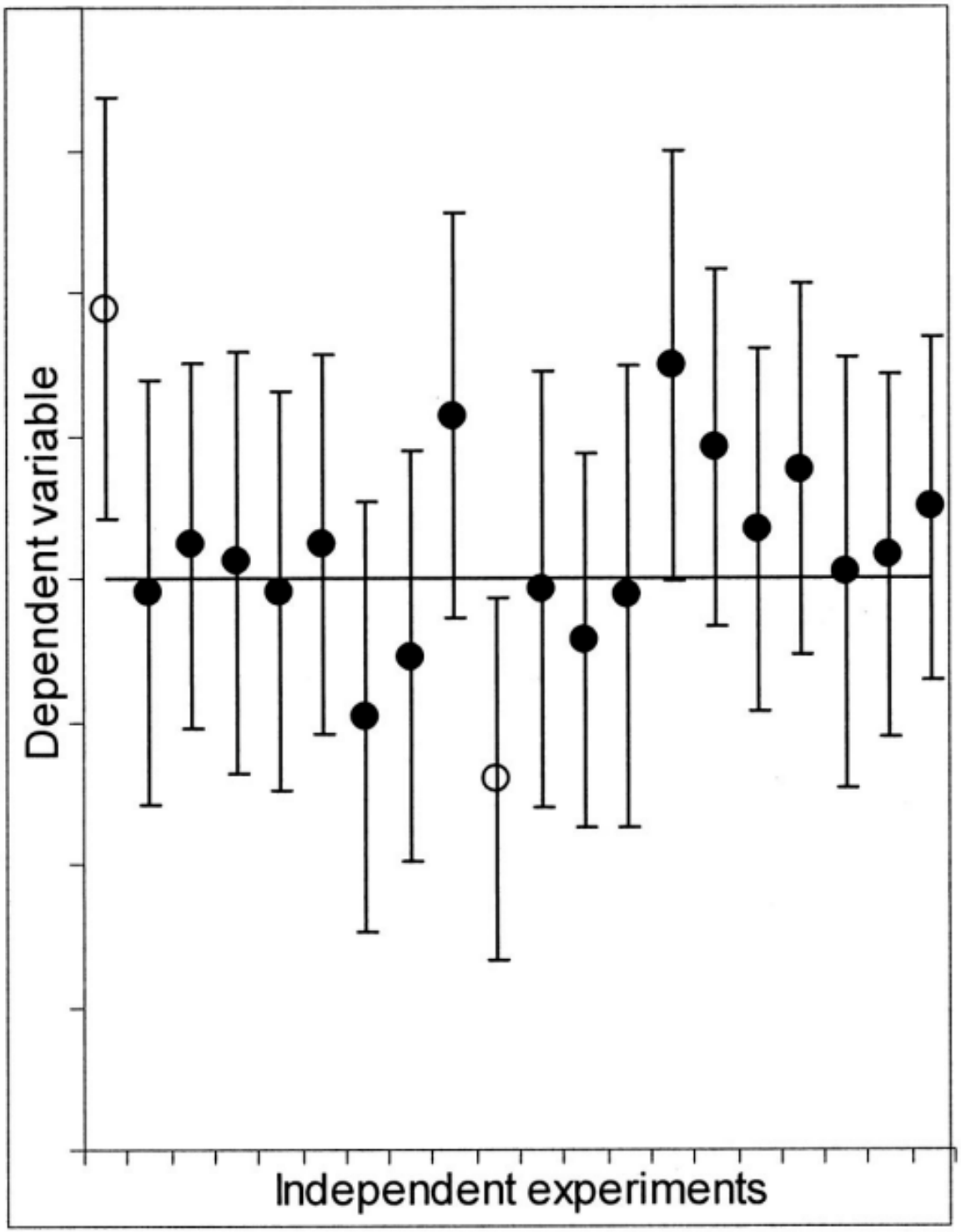
- « *a range of plausible values for μ. Values outside the CI are relatively implausible.* »
  (Cumming and Finch, 2005)



How to Interpret CIs?

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

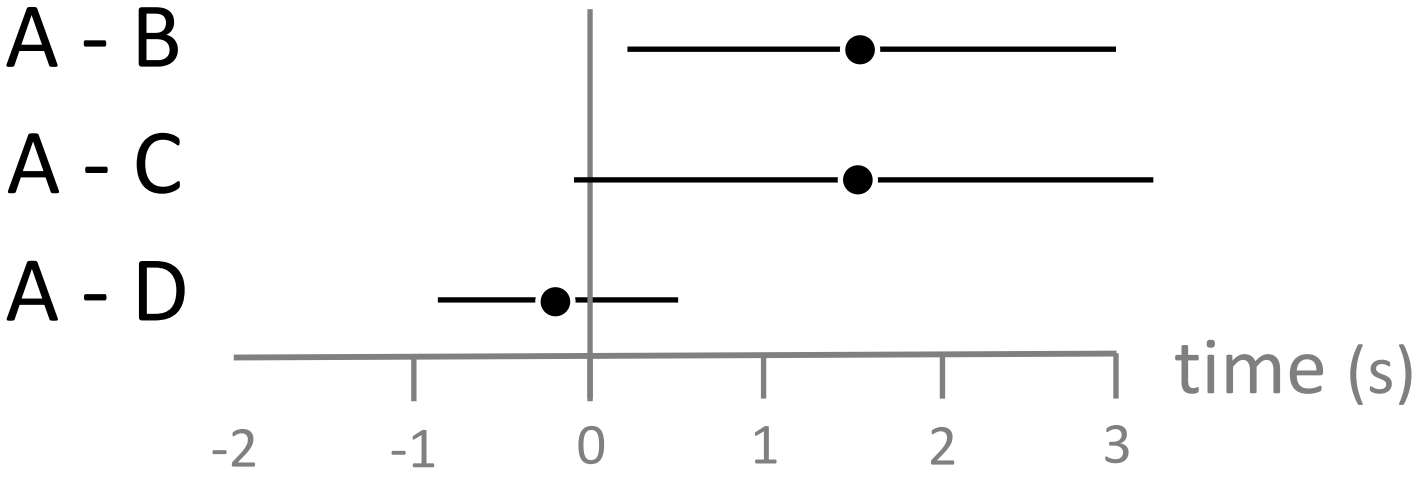




- Make sure you check *the dance of p-values* on youtube

How to Interpret CIs?

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



How to Interpret CIs?

“ It seems clear that no confidence interval should be interpreted as a a significance test.”

(Schmidt and Hunter, 1997)

How to Interpret CIs?

- Very hard!
  - We believe that a user study should provide yes/no answers
  - We believe that we need an objective procedure for deciding
  - We've been brainwashed!

How to Interpret CIs?

“ It is best for individual researchers to present point estimates and confidence intervals and ***refrain from attempting to draw final conclusions*** *about research hypotheses. ”*

Schmidt and Hunter (1997)

“ We have the duty of […] communicating our conclusions in intelligible form, in recognition of ***the right of other free minds*** *to utilize them in* ***making their own decisions****. ”*

Fisher (1955)

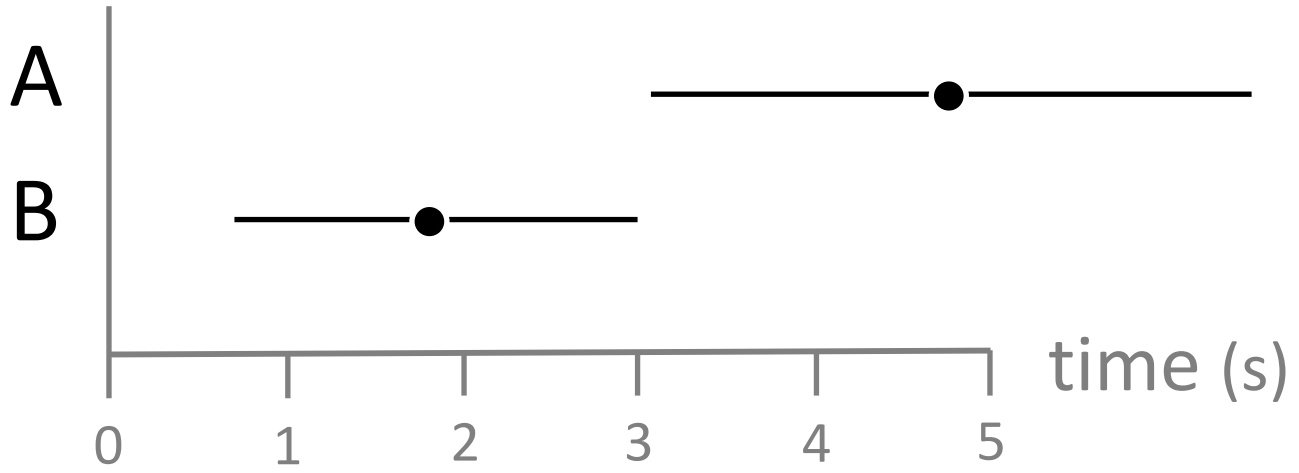
How to Interpret CIs?

"[…] (Sciences) can only be successfully conducted by responsible and independent thinkers […] The idea that this responsibility can be delegated to a giant computer programmed with Decision Functions belongs to the phantasy of circles rather remote from scientific research. "

Fisher (1973), quoted by Smith et al. (2002)

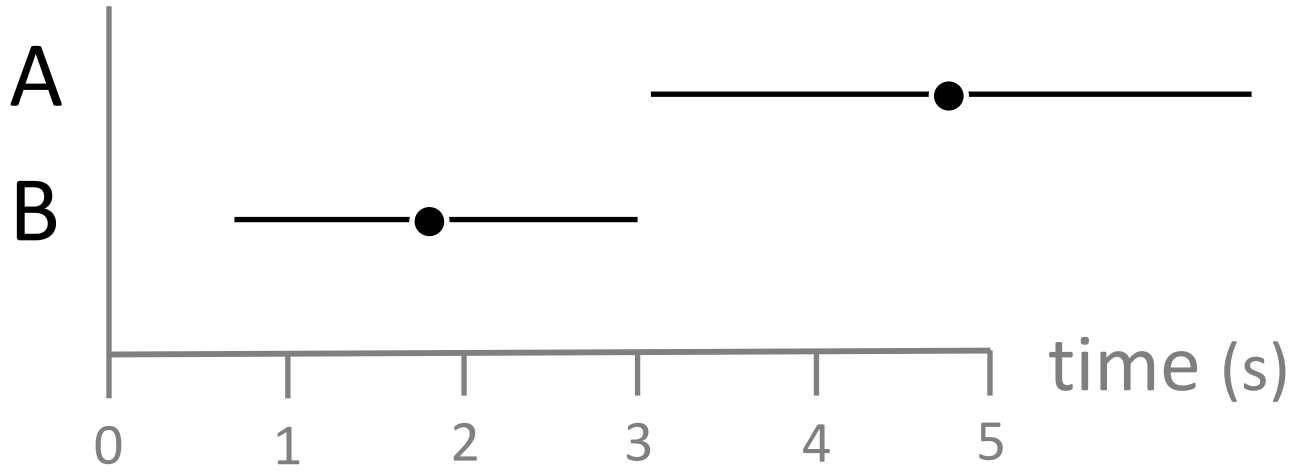
How to Interpret CIs?

- Overlap between CIs



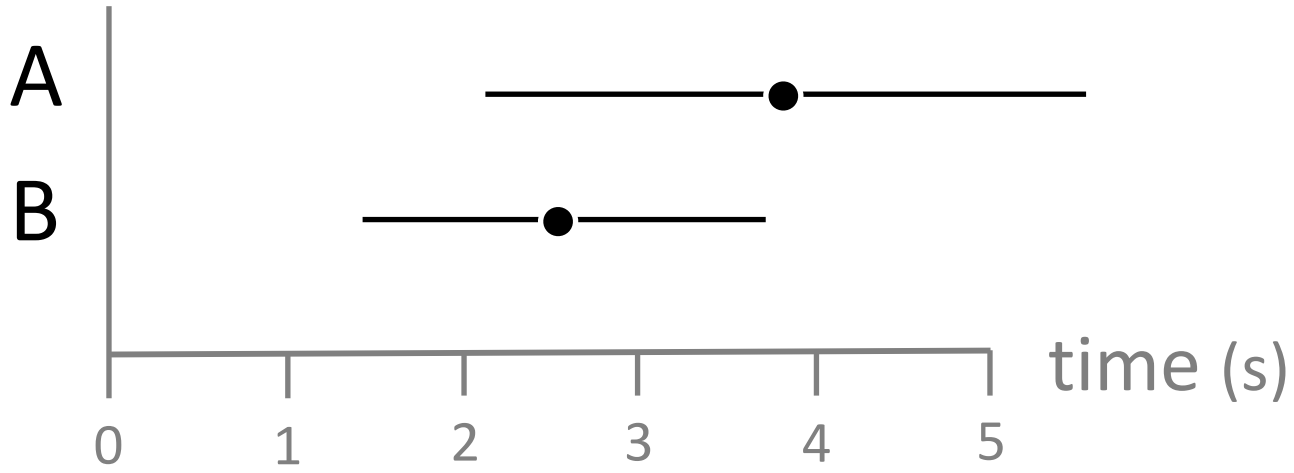
How to Interpret CIs?

- Overlap between CIs
  - Case of **between-subjects** design
  - Is the difference statistically significant?



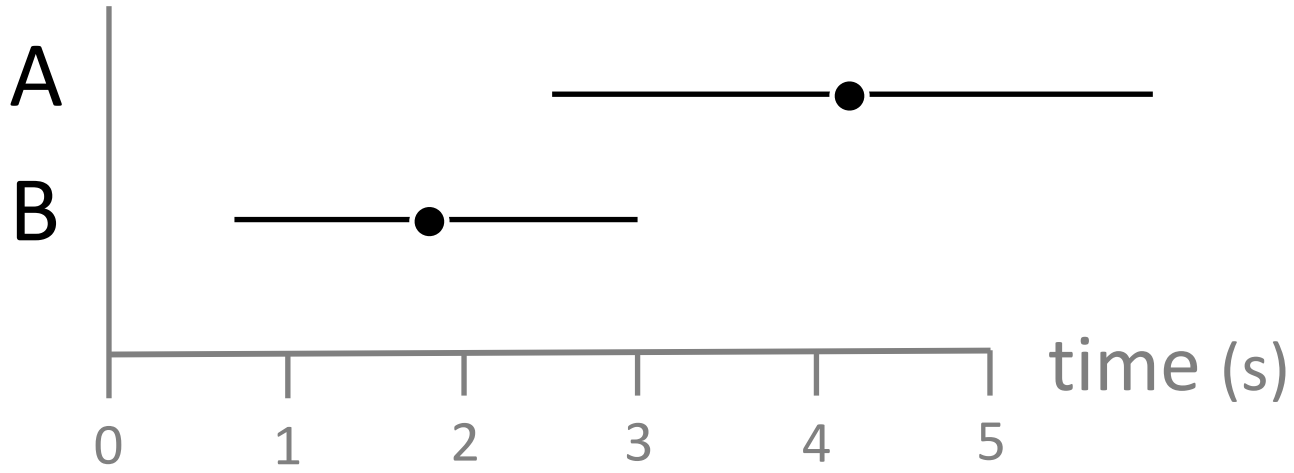
How to Interpret CIs?

- Overlap between CIs
  - Case of **between-subjects** design
  - Is the difference statistically significant?



How to Interpret CIs?

- Overlap between CIs
  - Case of **between-subjects** design
  - Is the difference statistically significant?

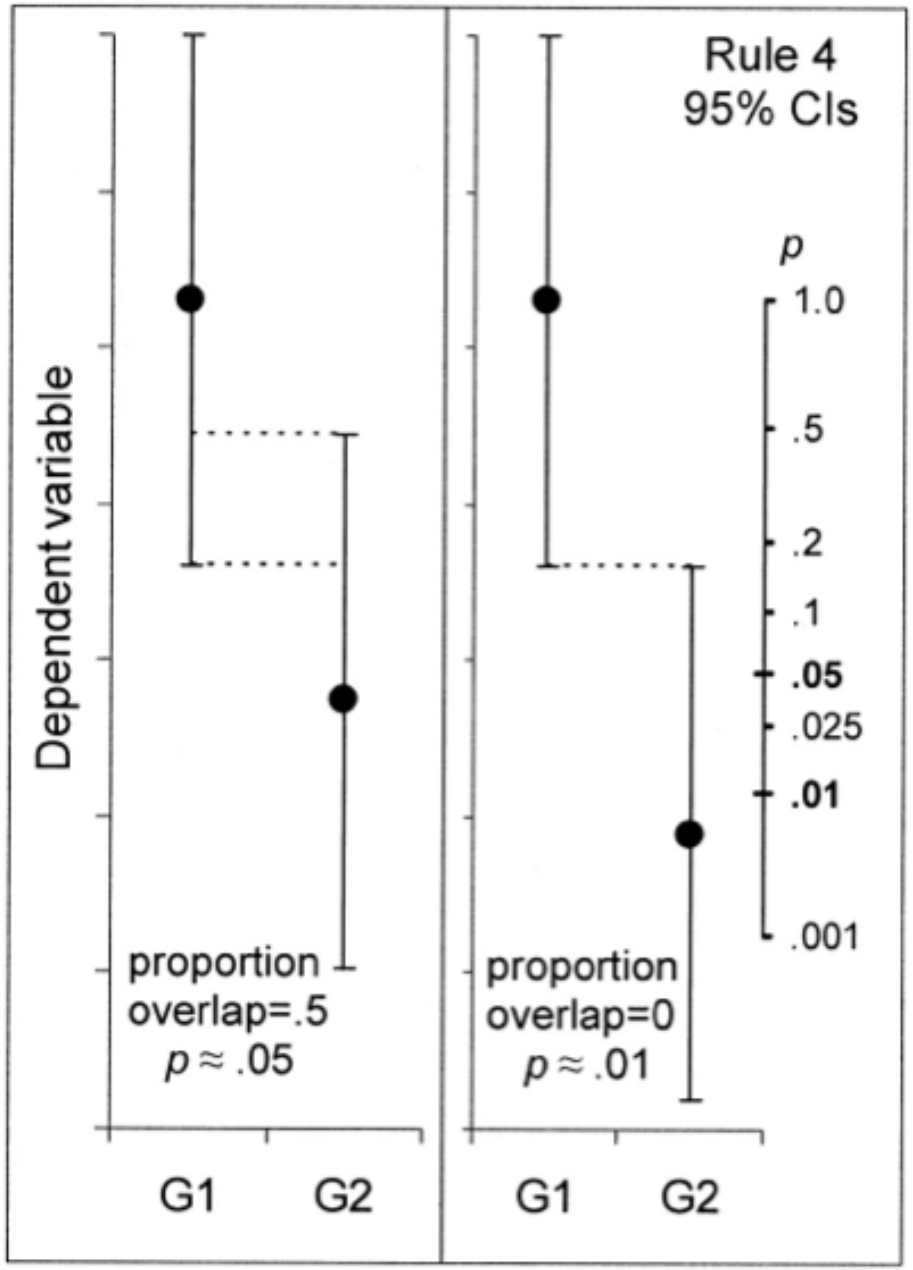


How to Interpret CIs?

- Overlap between CIs

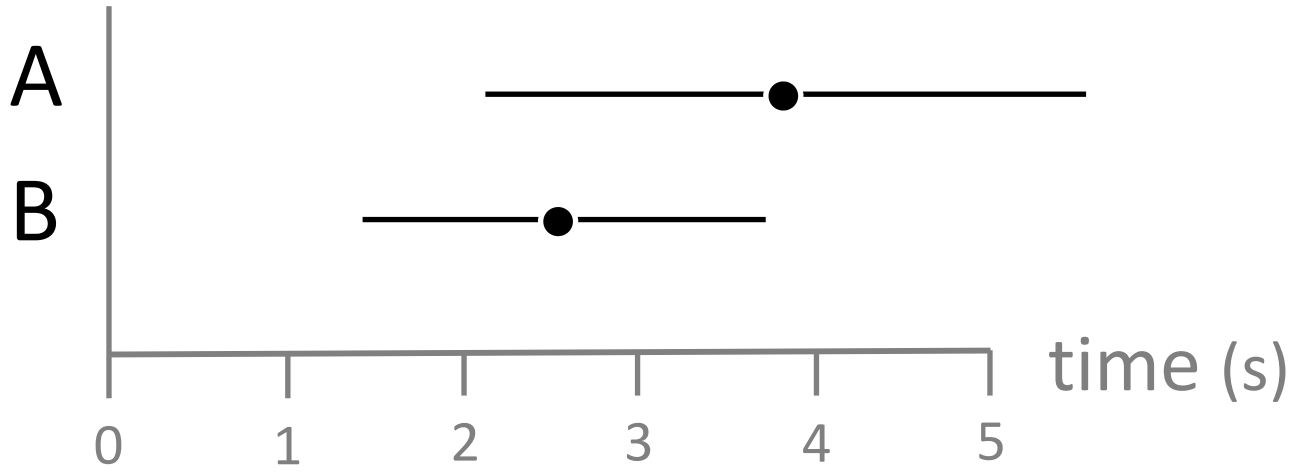
  Cumming and Finch's

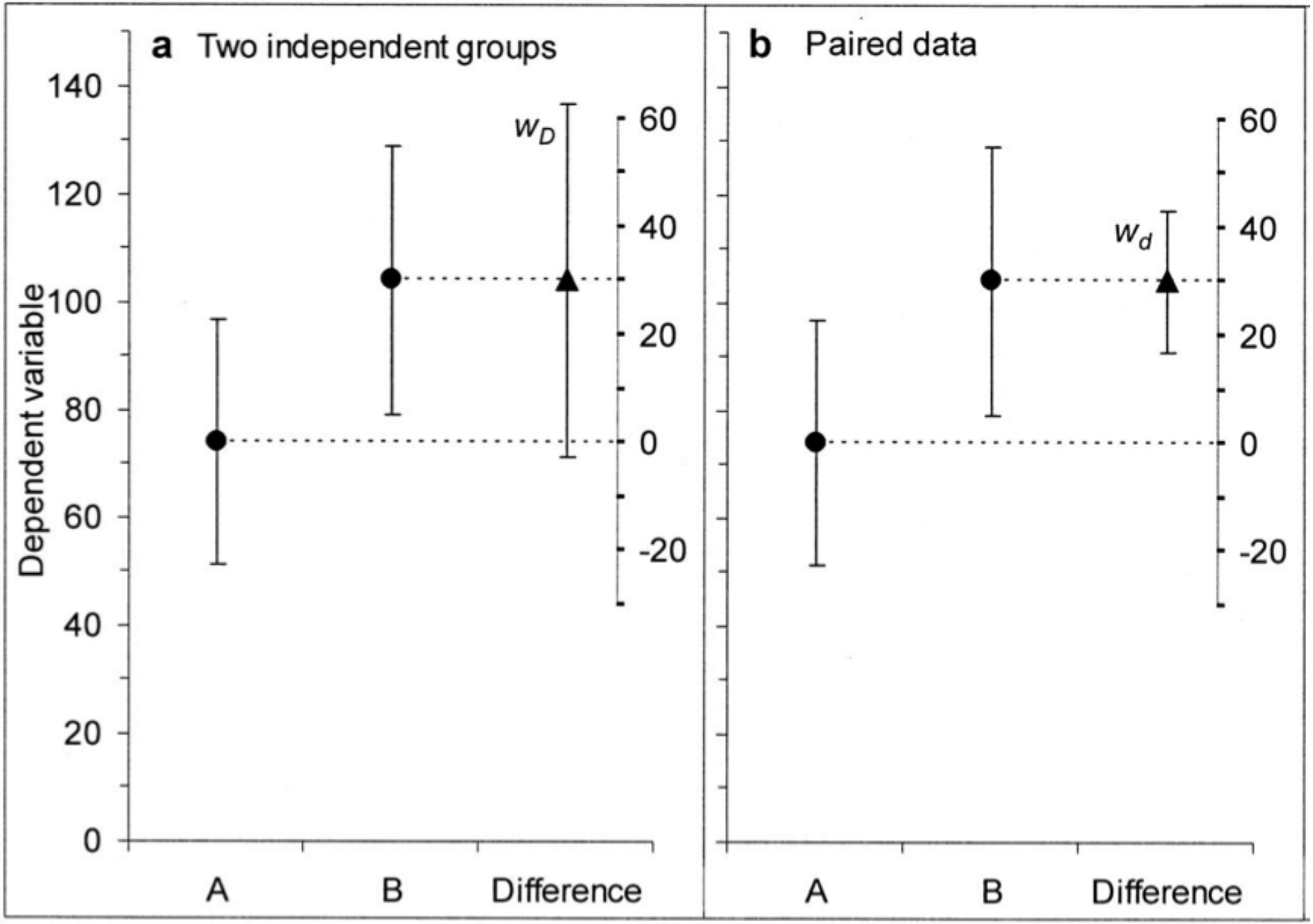
  Rule of Eye
  (Cumming and Finch, 2005)



How to Interpret CIs?

- Overlap between CIs
  - Case of **within-subject** design
  - Is the difference statistically significant?





(Cumming and Finch, 2005)

How to Interpret CIs?

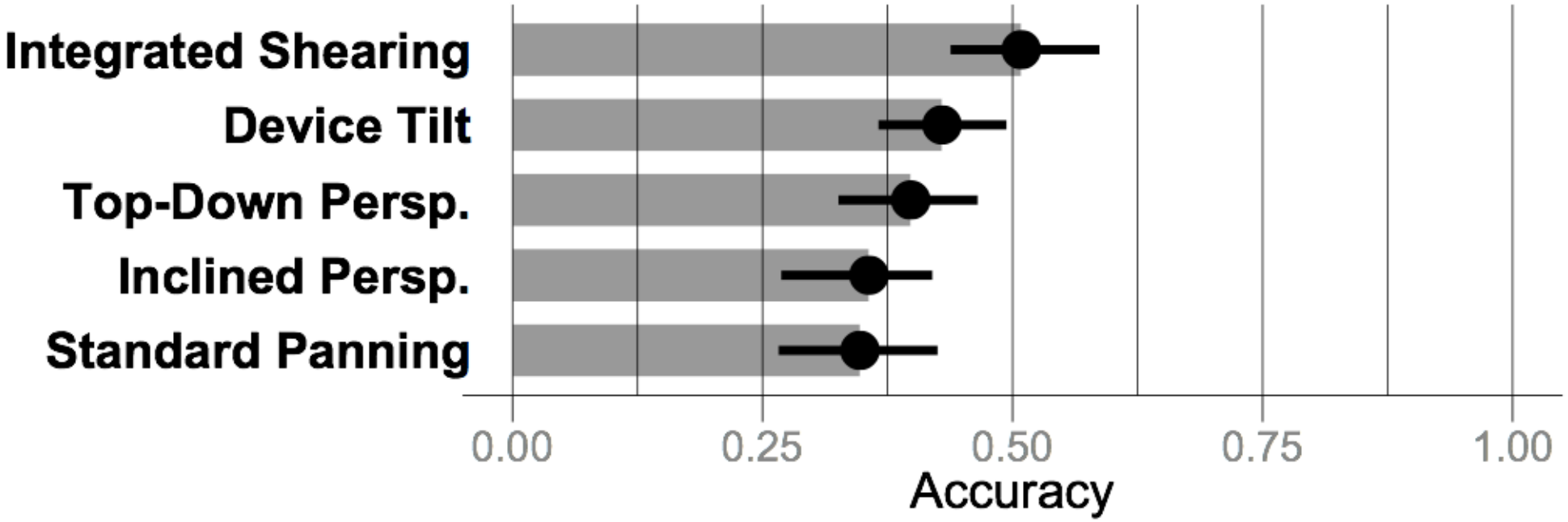
- What if there are several factors/levels?
  - Inferential (or ANOVA) confidence intervals
  - Corrections for multiple comparisons
  - Complicated to compute AND to interpret

How to Interpret CIs?

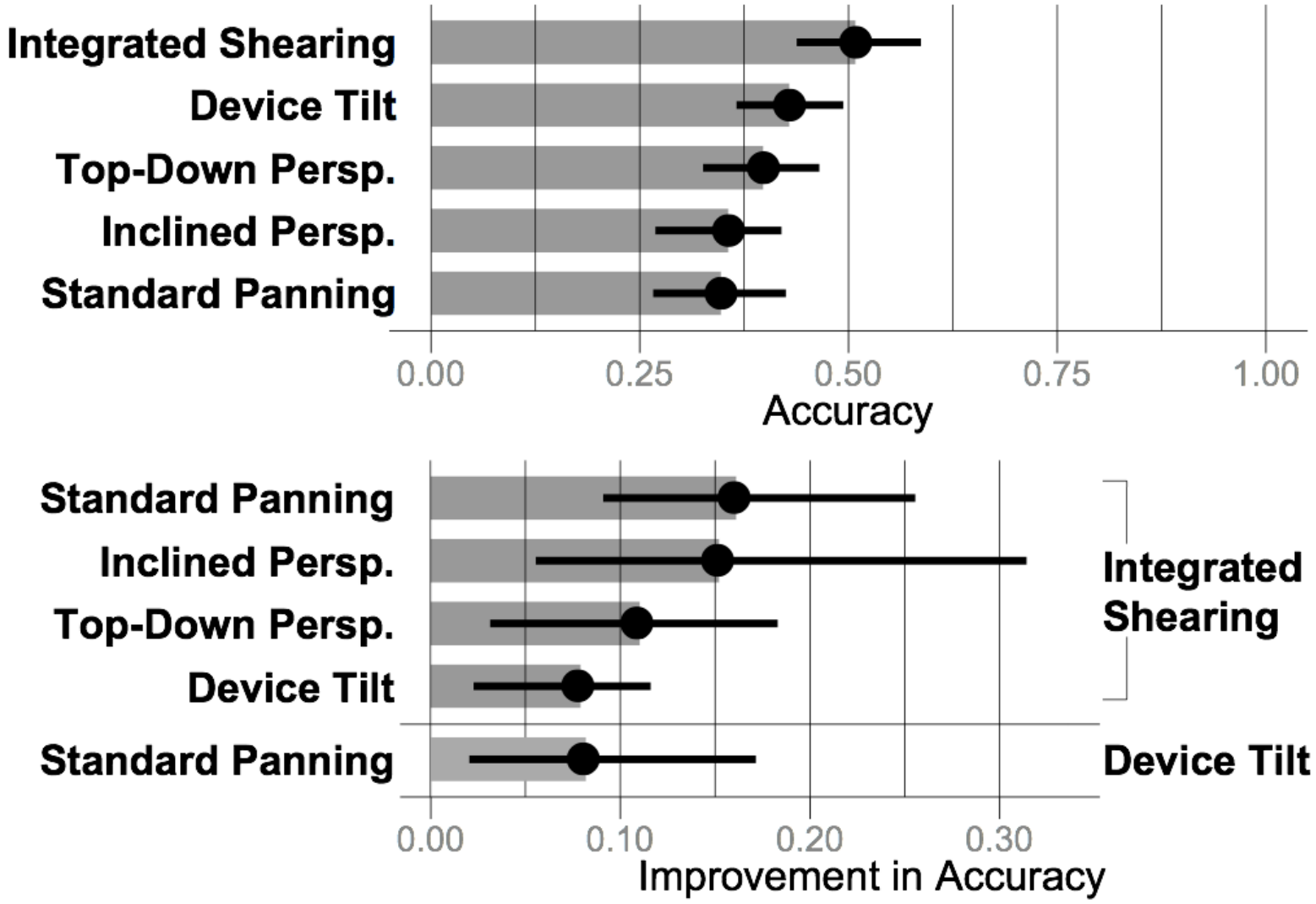
- What if there are several factors/levels?
  - Choose a simple experiment design
  - Pre-specify your research questions in advance
  - Only show and interpret the effects of interest
  - Don’t correct for multiple comparisons
  - Do all your analyses on pilot data FIRST

Some Real Examples

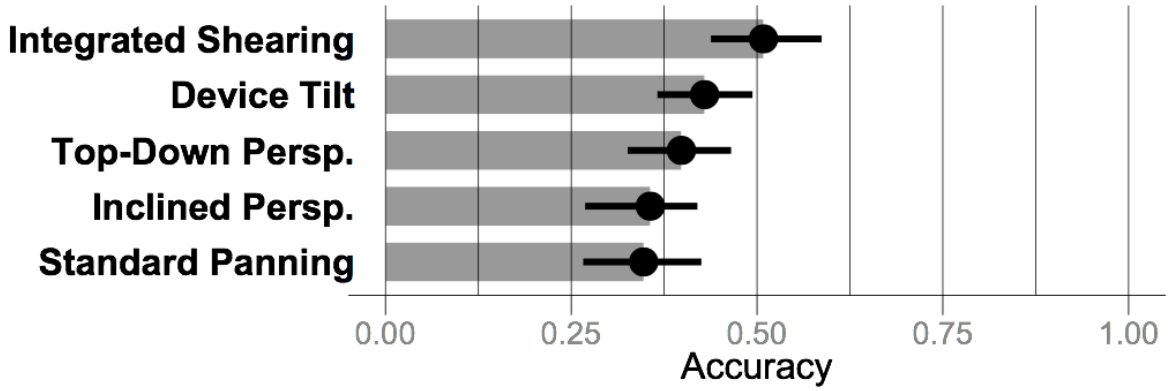
1 within-subject factor: *technique* (5 levels)
2 measures: *accuracy* and *time*

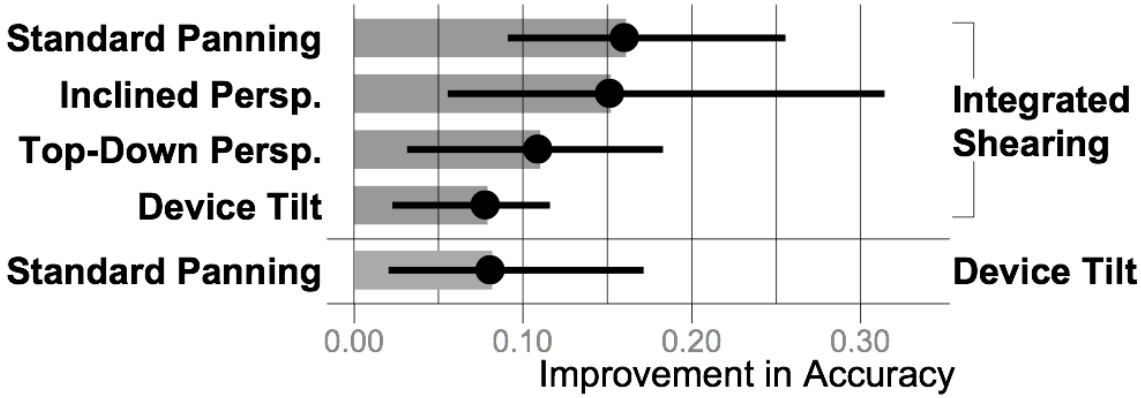


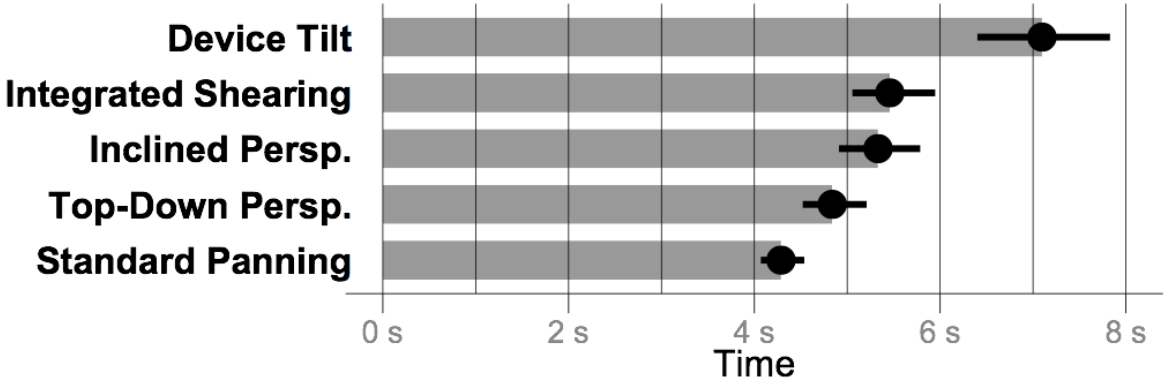
1 within-subject factor: *technique* (5 levels)
2 measures: *accuracy* and *time*



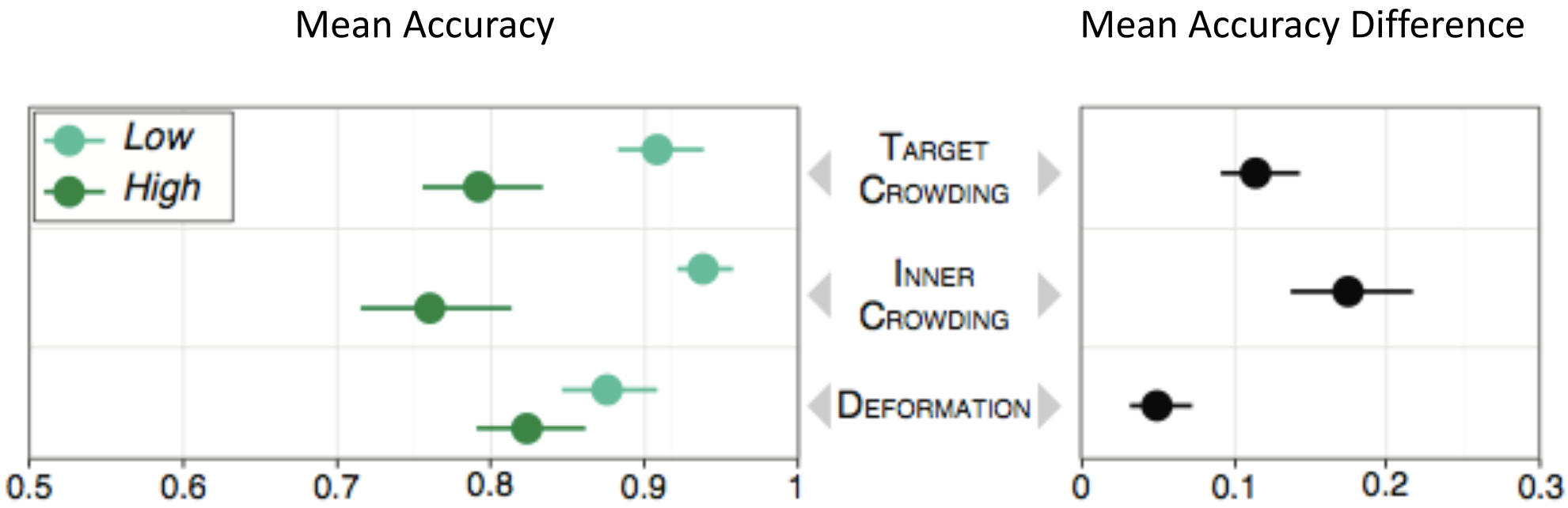
1 within-subject factor: *technique* (5 levels)
2 measures: *accuracy* and *time*



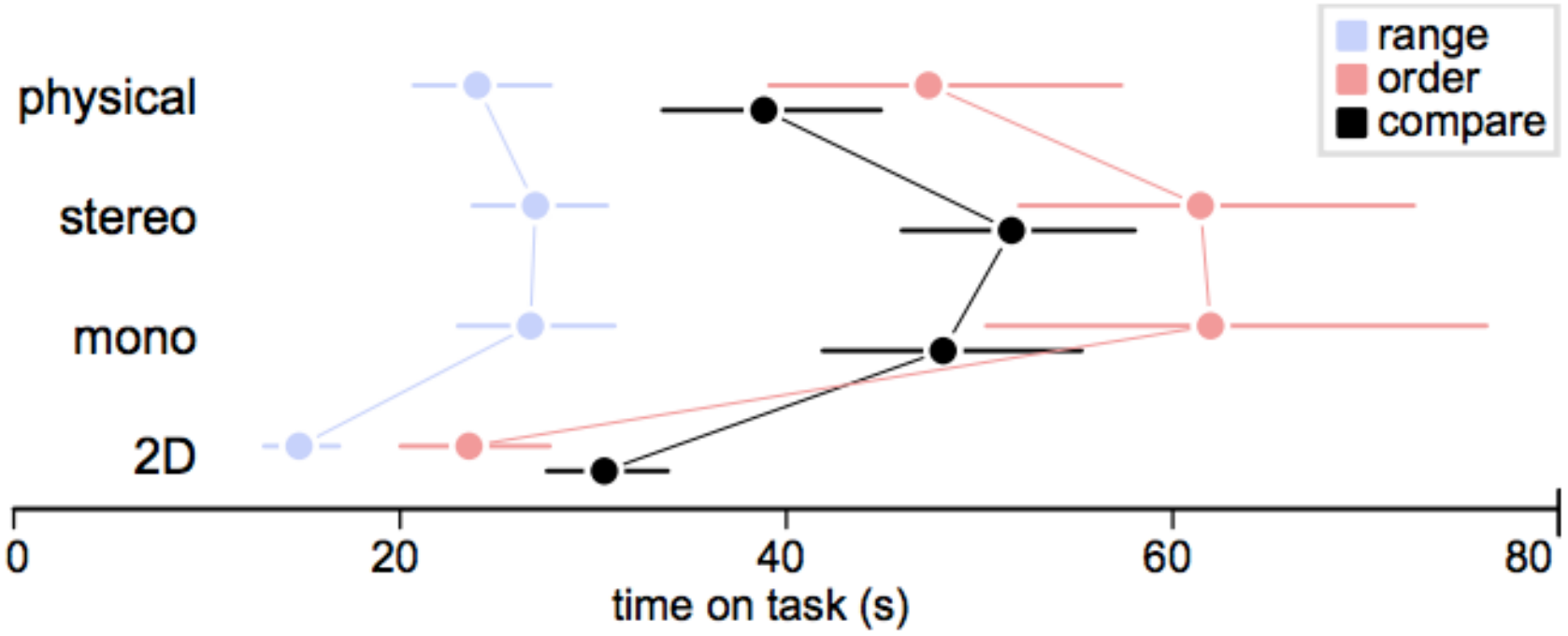




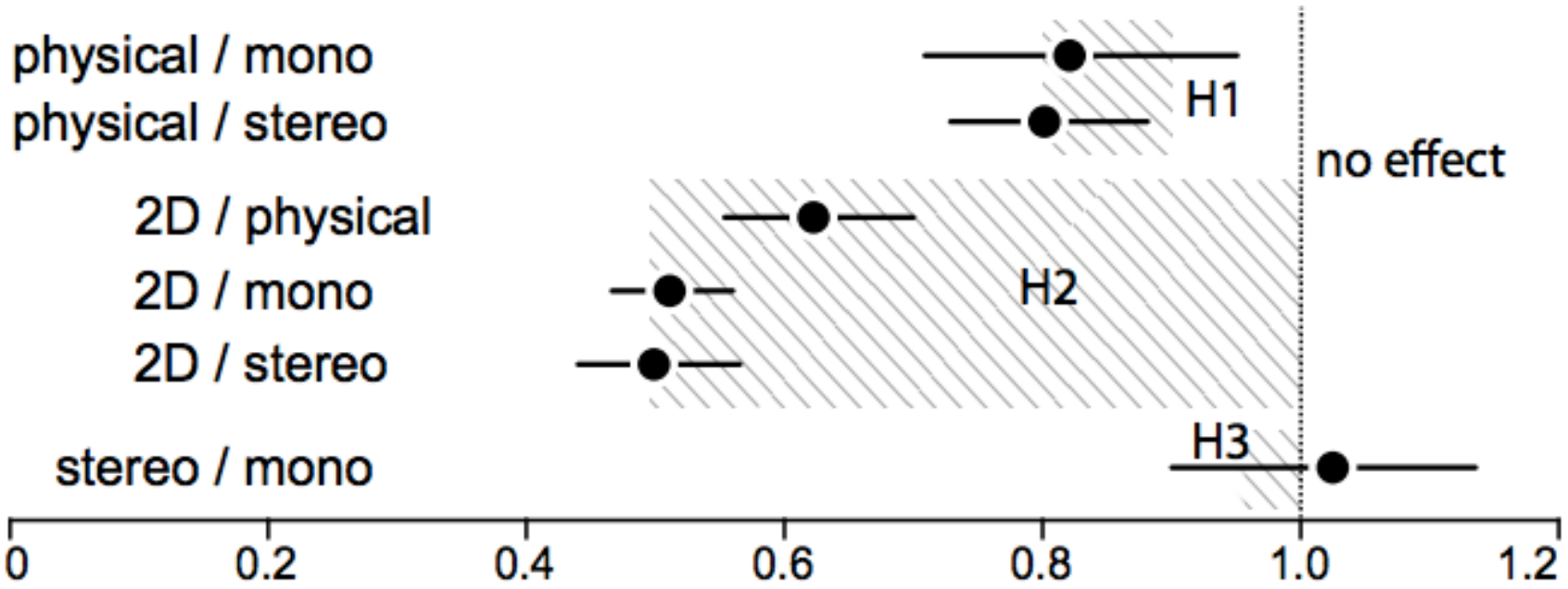
3 within-subject factors: *target crowding*, *inner crowding*, *deformation* (2 levels each)
1 measure: *accuracy*



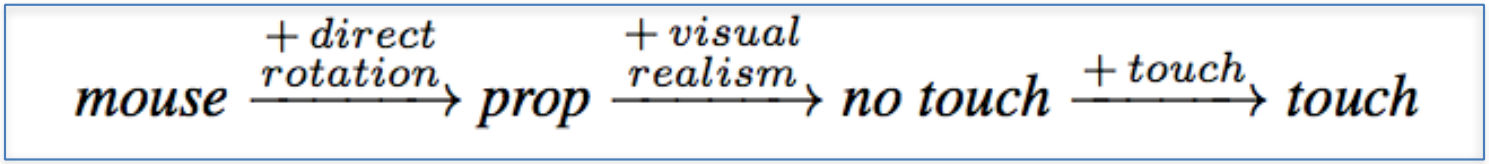
2 within-subject factors: *technique* (4 levels) and *task* (3 levels) 1 measure: *time*

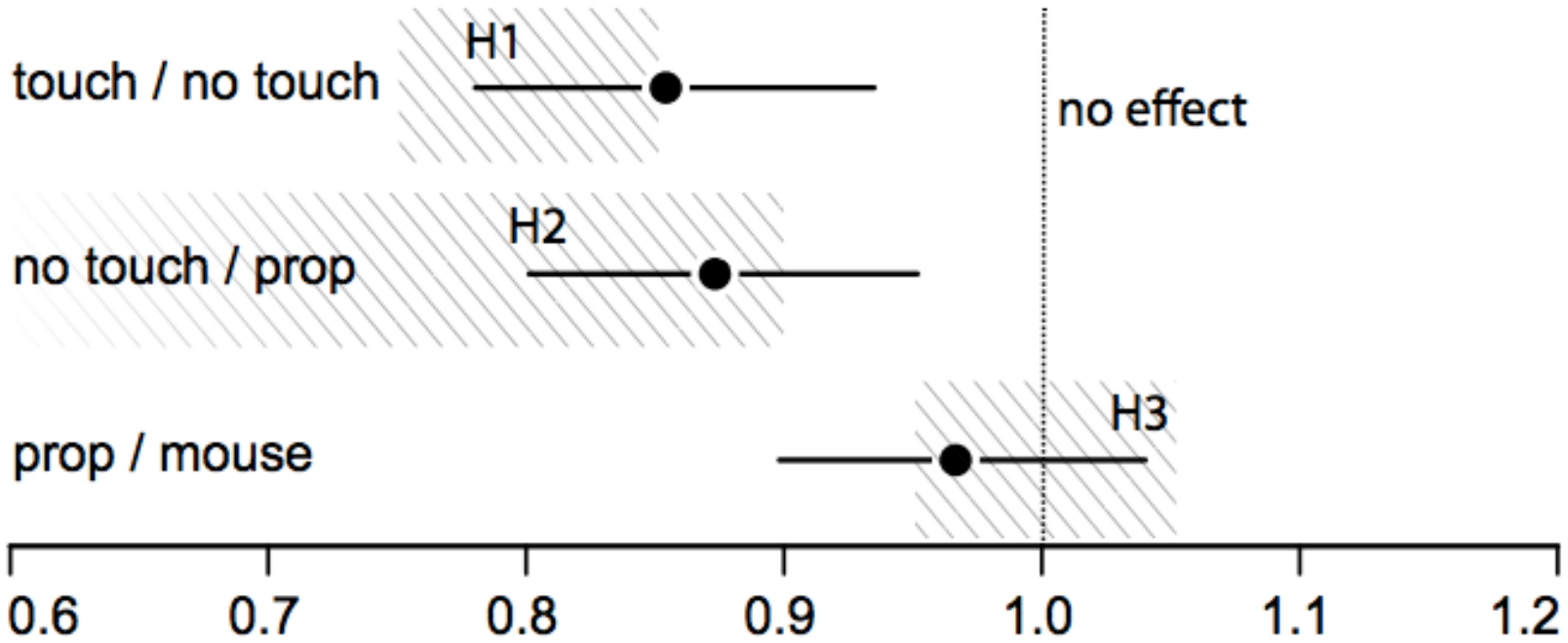


2 within-subject factors:
technique (4 levels) and *task* (3 levels)
1 measure: *time*



1 within-subject factor: *technique* (4 levels)
1 measure: *time*





Will My Paper be Rejected?

- No (most likely)
  - If you don't over-interpret the patterns in your CIs
  - If you properly justify your approach

Due to growing concerns in various research fields over the limits of null hypothesis significance testing for reporting and interpreting experimental results [12], we base all our analyses and discussions on estimation, i.e., effect sizes with confidence intervals [13]. This approach also aligns with the latest recommendations from the APA [3].

To Go Further

- **Geoff Cumming**
  - Youtube channel
  - Book: "The New Statistics"
- **Allen Downey**
  - Book "Think Stats - Probability and Statistics for Programmers" (also a lecture)
- www.aviz.fr/badstats
  - Reading list on the *p*-value controversy
  - Examples of HCI papers without *p*-values