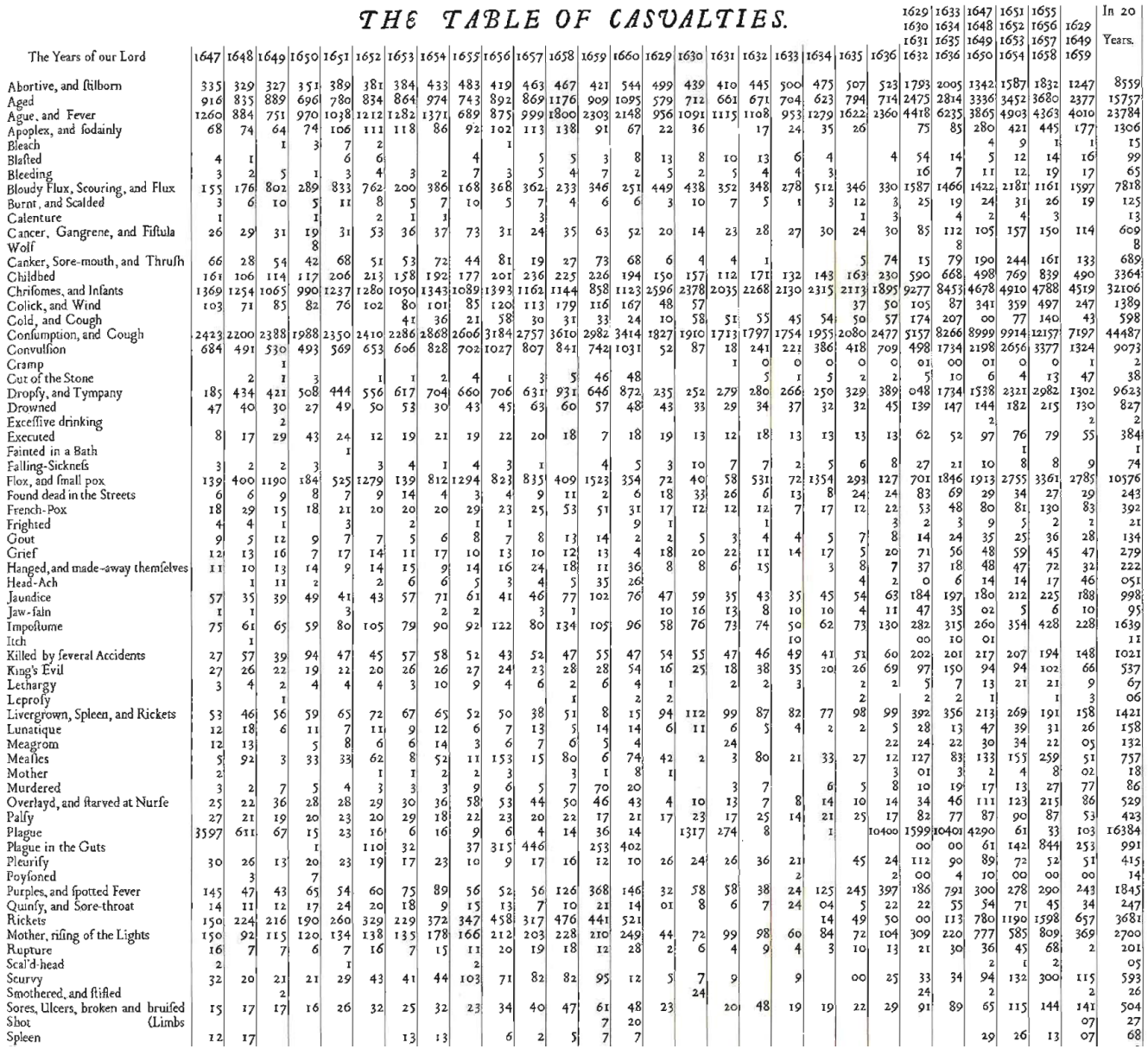
Introduction to data science & artificial intelligence (INF7100)

Arthur Charpentier

#221 Statistical Inference: Average

été 2020

Statistical Indicators



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	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	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	668	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	80	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	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	048	1734	1538	2321	2982	1302	9623
Drowned	47	40	30	27	49	50	53	30	43	45	63	60	57	48	43	33	29	34	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	1846	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	26	6	13	8	24	24	83	69	29	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	56	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
Head-Ach		1	11	2		2	6	6	5	3	4	5	35	26							4	2	0	6	14	14	17	46	051
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	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	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	158
Meagrom	12	13		5	8	6	6	14	3	6	7	6	5	4			24					22	24	22	30	34	22	05	132
Meaſles	5	92	3	33	33	62	8	52	11	153	15	80	6	74	42	2	3	80	21	33	27	12	127	83	133	155	259	51	757
Mother	2					1	1	2	2	3		3	1	8	1							3	01	3	2	4	8	02	18
Murdered	3	2	7	5	4	3	3	3	9	6	5	7	70	20			3	7		6	5	8	10	19	17	13	27	77	86
Overlayd, and ſtarved at Nurſe	25	22	36	28	28	29	30	36	58	53	44	50	46	43	4	10	13	7	8	14	10	14	34	46	111	123	215	86	529
Palſy	27	21	19	20	23	20	29	18	22	23	20	22	17	21	17	23	17	25	14	21	25	17	82	77	87	90	87	53	423
Plague	3597	611	67	15	23	16	6	16	9	6	4	14	36	14		1317	274	8		1		10400	1599	10401	4290	61	33	103	16384
Plague in the Guts				1		110	32		37	315	446		253	402									00	00	61	142	844	253	991
Pleuriſy	30	26	13	20	23	19	17	23	10	9	17	16	12	10	26	24	26	36	21		45	24	112	90	89	72	52	51	415
Poyſoned		3		7															2			2	00	4	10	00	00	00	14
Purples, and ſpotted Fever	145	47	43	65	54	60	75	89	56	52	56	126	368	146	32	58	58	38	24	125	245	397	186	791	300	278	290	243	1845
Quinſy, and Sore-throat	14	11	12	17	24	20	18	9	15	13	7	10	21	14	01	8	6	7	24	04	5	22	22	55	54	71	45	34	247
Rickets	150	224	216	190	260	329	229	372	347	458	317	476	441	521						14	49	50	00	113	780	1190	1598	657	3681
Mother, riſing of the Lights	150	92	115	120	134	138	135	178	166	212	203	228	210	249	44	72	99	98	60	84	72	104	309	220	777	585	809	369	2700
Rupture	16	7	7	6	7	16	7	15	11	20	19	18	12	28	2	6	4	9	4	3	10	13	21	30	36	45	68	2	201
Scal'd-head	2				1				2																2	[illegible]	2		05
Scurvy	32	20	21	21	29	43	41	44	103	71	82	82	95	12	5	7	9		9		00	25	33	34	94	132	300	115	593
Smothered, and ſtifled			2													24							24		2			2	26
Sores, Ulcers, broken and bruiſed	15	17	17	16	26	32	25	32	23	34	40	47	61	48	23		20	48	19	19	22	29	91	89	65	115	144	141	504
Shot (Limbs													7	20														07	27
Spleen	12	17					13	13		6	2	5	7	7											29	26	13	07	68

Average, mean, median, mode, etc

Given a sample $\boldsymbol{x} = \{x_1, \cdots, x_n\}$, the average is $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Note that $\overline{x} = \underset{m \in \mathbb{R}}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} (x_i - m)^2 \right\}$

In python

```
> import statistics
> x = [1, 2, 3, 4, 5, 6]
> statistics.mean(x)
3.5
```

and in R

```
> x = 1:6
> mean(x)
[1] 3.5
```

Average, mean, median, mode, etc

The average (mean) is the empirical version of the expected value of a random variable,

$$\mathbb{E}(X) = \sum_x x\mathbb{P}[X = x] \text{ or } \int xf(x)dx$$

Example: a coin has *heads* with probability p. Let $x = \mathbf{1}(heads)$,

$$\mathbb{E}(X) = 1 \cdot p + 0 \cdot (1 - p) = p$$

Linearity:

$$\mathbb{E}(aX + b) = a\mathbb{E}(X) + b, \ \forall a, b \in \mathbb{R}, X$$

$$\mathbb{E}(X_1 + \cdots + X_k) = \mathbb{E}(X_1) + \cdots \mathbb{E}(X_k), \ \forall X_1, \cdots, X_k$$

Example: toss n coins, of bias p, X is the number of heads

$$\mathbb{E}(X) = \mathbb{E}(X_1 + \cdots + X_n) = \mathbb{E}(X_1) + \cdots \mathbb{E}(X_n) = np$$

St Petersburg's Paradox

As we will see (law of large numbers) if x_i are realizations of random variables X_i (with identical expected value μ), $\overline{x} \to \mu$ as $n \to \infty$.

A fair coin is tossed at each stage. The initial stake begins at 2 dollars and is doubled every time heads appears. The first time tails appears, the game ends and the player wins whatever is in the pot. Let X denote the gain.

$$\mathbb{E}(X) = \frac{1}{2} \cdot 2 + \frac{1}{4} \cdot 4 + \frac{1}{8} \cdot 8 + \frac{1}{16} \cdot 16 + \cdots = 1 + 1 + 1 + 1 + \cdots = +\infty$$

the expected value is infinite (but the average always exists)

Average, mean, median, mode, etc

The average is very sensitive to outliers and extremal values.
Denote $\{x_{(i)}\}$ the ordered version of $\{x_i\}$, $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$

Trimmed mean $= \dfrac{1}{n-2p} \displaystyle\sum_{i=1+p}^{n-p} x_{(i)}$

Weighted mean $= \displaystyle\sum_{i=1}^{n} \omega_i x_i$ where $\omega_i = \dfrac{w_i}{w_1 + \cdots + w_n}$

(classical on surveys with underrepresentation, see fao or Stat Can)

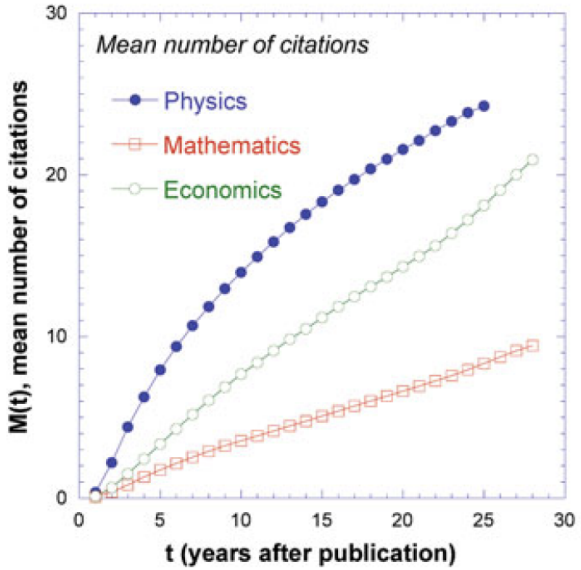
Median $md(x) = x_{(n/2)}$ (50% observations are smaller/larger)

Note that $md(x) = \underset{m \in \mathbb{R}}{\text{argmin}} \left\{ \displaystyle\sum_{i=1}^{n} |x_i - m| \right\}$ (see #224)

Average ratio

See Citation Analysis and Dynamics of Citation Networks by Michael Golosovsky.

	0	1	2	3	4	5
2014	122	146	163	176	182	186
2015	142	174	198	214	228	
2016	185	215	246	278		
2017	214	265	312			
2018	245	326				
2019	261					

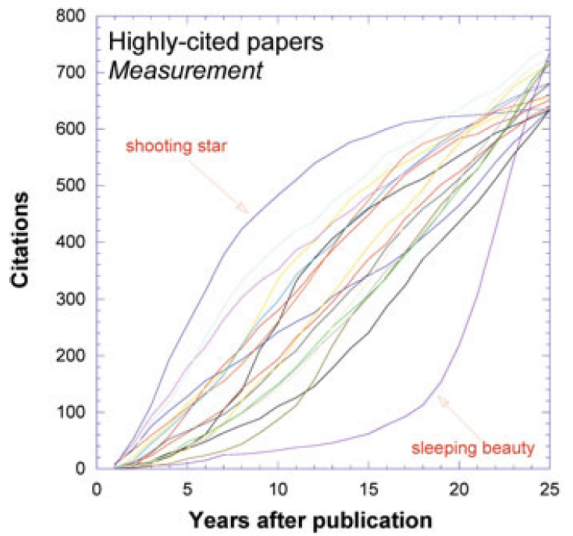


The average ratio is

$$\frac{326+265+215+174+146}{245+214+185+142+122}$$

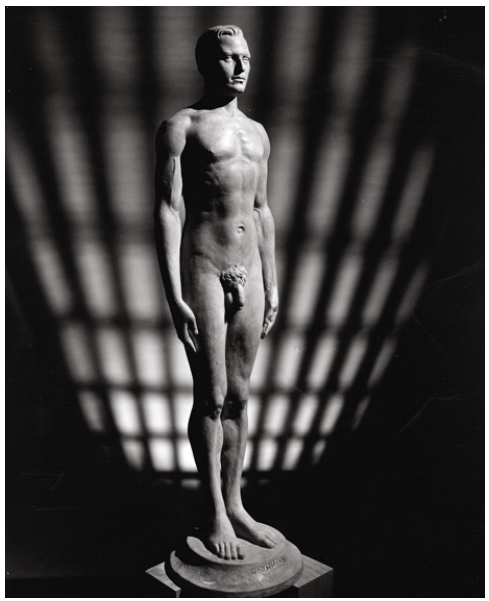
and <u>not</u> the average of ratios

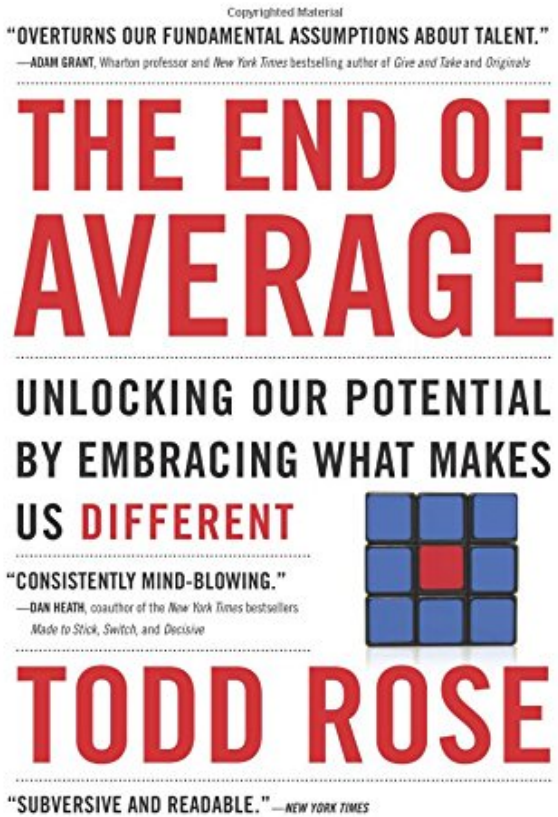
$$\frac{1}{5}\left(\frac{326}{245}+\frac{265}{214}+\frac{215}{185}+\frac{174}{142}+\frac{146}{122}\right)$$



Norms and Averages

See The End of Average, by Todd Rose



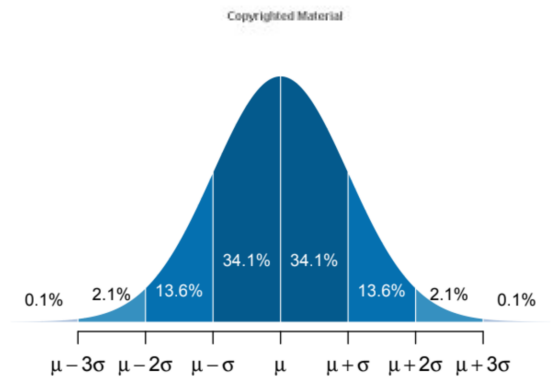


$$\mathbb{P}(Z_1 \in [\pm 2\sigma]) = 95\%, 5\% \text{ outliers}$$

$$\mathbb{P}(Z_1, Z_2 \in [\pm 2\sigma]) = 91\%, 9\% \text{ outliers}$$

$$\mathbb{P}(Z_1, Z_2, \cdots, Z_{15} \in [\pm 2\sigma]) = 50\%, 50\% \text{ outliers}$$

There is no average person in high dimension (see also curse of dimensionality)



What is an Average

"*Take any stock in the United States. The average time in which you hold a stock is – it's gone up from 20 seconds to 22 seconds in the last year*" The Telegraph, by Michael Hudson

"*Even though most are judged by performance over three-year horizons, their average holding period was about 17 months, and 19% of the managers held the typical stock for one year or less*" The Wall Street Journal, by Jason Zweig