Written 11/5/2013
Updated 5/7/2017

Twyman's Law and Controlled Experiments

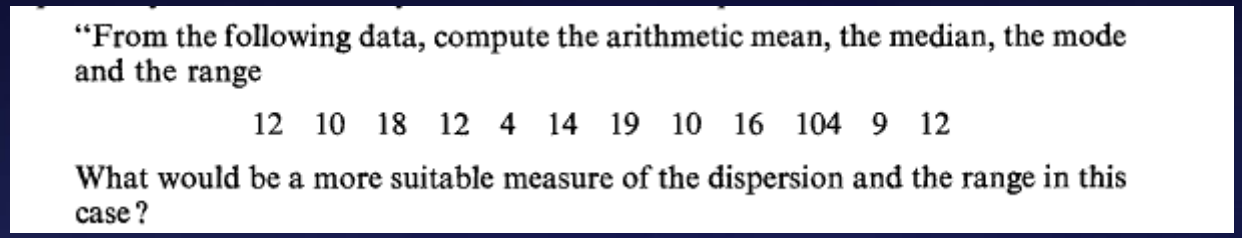
http://bit.ly/twymanLaw

Ronny Kohavi

Twyman's Law

Any figure that looks interesting or different is usually wrong

- Earliest scholarly reference I found is in Journal of the Royal Statistical Society, Series A, Vol 138, No 4, 1975.
  The Teaching of Statistics by A. S. C. Ehrenberg



"From the following data, compute the arithmetic mean, the median, the mode and the range

12 10 18 12 4 14 19 10 16 104 9 12

What would be a more suitable measure of the dispersion and the range in this case?

On the one hand, the student is presumably meant to query the 104. Should it have been 10 or 14, or perhaps 10 *and* 4? (Twyman's Law: "Any figure that looks interesting or different is usually wrong.")

- I recall an emetrics talk with the law stated as
  Any statistic that appears interesting is almost certainly a mistake
  and I used it in an invited talk (Focus the Mining Beacon in 2005)

Tracing Twyman's Law

- William Anthony Twyman, W. A. Twyman, and Tony Twyman refer to the UK radio and television audience measurement veteran
- He died Oct 31, 2014, at the age of 82 (Obituary)
- From what I can tell, he never published the law attributed to him
- Andrew S. C. Ehrenberg cited Twyman's law in Teaching of Statistics (prior slide) and in Rudiments of Numeracy as

  Any reading which looks interesting or different is probably wrong
- Ehrenberg and Twyman published On measuring television audiences in 1967, but that paper does not mention the "law."
- Official Rules and Explanations cite the rule as (see prior slide):

  Any statistic that appears interesting is almost certainly a mistake

  It is "From Iwan Williams, London, who got it from A.S.C. Ehrenberg, Professor of Marketing, who got it from a colleague."

Other References to Twyman's law

- Exploring Data: An Introduction to Data Analysis for Social Scientists by Catherine Marsh, Jane Elliott wrote

  > Twyman's law: The more unusual or interesting the data, the more likely they are to have been the result of an error of one kind or another

  The authors claim that Twyman's law is "perhaps the most important single law in the whole of data analysis."
- Statistically Speaking: A Dictionary of Quotations by C.C. Gaither, Alma E Cavazos-Gaither [1996] claims the origin is unknown
- Sampling in Archaeology by Clive Orton [2000] and Statistics, an Appraisal both site Ehrenberg's 1975 article

Examples

- Here are some examples from my experiences
  - If you have a mandatory birth date field and people think it's unnecessary, you'll find lots of people born on 11/11/11 or 01/01/01
  - If you have an optional drop down, do not default to the first alphabetical entry, or you'll have lots of: jobs = Astronaut
  - For most web sites, traffic had no traffic between 2AM and 3AM March 9, 2014. Don't worry, there was no outage. It's daylight saving.
  - Multiple cases of Simpson's paradox

Common Mistakes

- One of the most common mistakes I see is that someone gets a surprising result and “explains” it with the wrong reason (e.g., this massive change is due to small perf)
- Another mistake is to focus on one of several changes as if it’s responsible for everything.
  - When you make three changes and succeed, separate them out
  - It’s possible they interact, but it’s more likely there’s a main factor
- The example on the next slide shows that when we don’t understand something, we can misinterpret a great result

Great Result Not Understood – Vitamin C

- Scurvy is a disease that results from vitamin C deficiency
- It killed over 100,000 people in the 16th-18th centuries, mostly sailors
- First known controlled experiment in 1747
  - Dr. James Lind noticed lack of scurvy in Mediterranean ships
  - Gave some sailors limes (treatment), others ate regular diet (control)
  - Experiment was highly successful
- But Lind didn't understand the reason
  - At the Royal Naval Hospital in England, he treated Scurvy patients with concentrated lemon juice called "rob."
  - He concentrated the lemon juice by heating it, thus destroying the vitamin C
  - He lost faith in the remedy and became increasingly reliant on bloodletting

Perf is Critical, but Not This Much

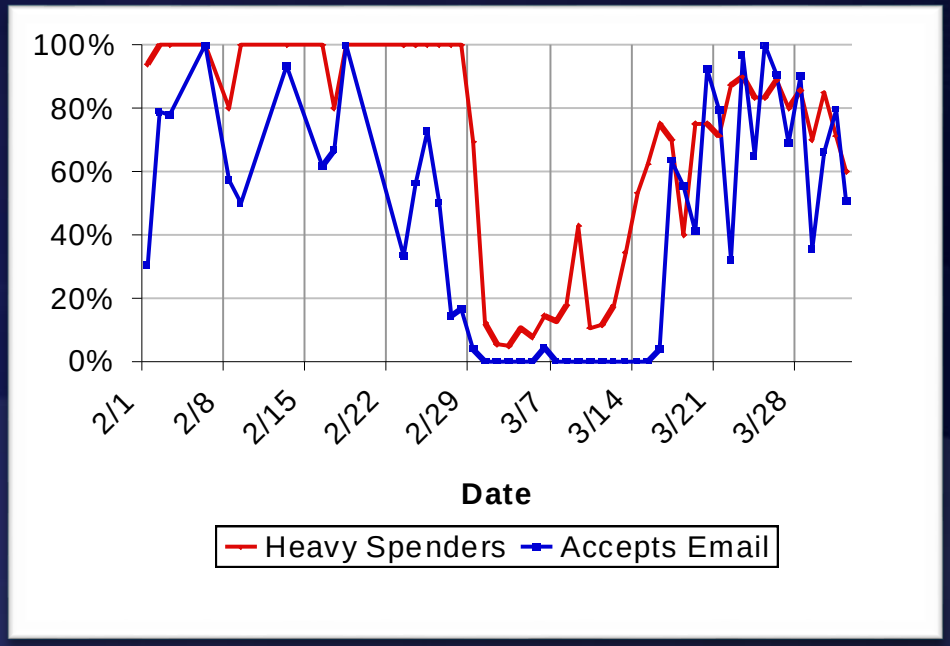
- At Web 2.0 in 2006, there was a great story of how Google tested increasing the number of results from 10 to 20
  - Revenue reduced by 20%
  - Explanation? time to display results increased 500msecs
- Twyman's law in action: perf is critical, but 500msec could explain only part of the decline
  - Our slowdown experiments (paper) show 100msec = 0.6% revenue impact. 500msec would be 3%, not 20%
  - We replicated this experiment and the revenue loss is likely the result of having a small portion of the page with ads.
    Instead of x mainline ads with 10 algorithmic results, you have x mainline ads with 20 algorithmic results, reducing ad prominence
  - We nullified the revenue loss by adding a single mainline ad (which slowed the page a bit more)
- See Seven Rules of Thumb for Web Site Experimenters, rule #3

Unreasonable Abandonment

- GeekWire's article from 10/29/2013: http://bit.ly/rapidAB
  - Google added another sponsored link
  - 30 day A/B test showed massive revenue improvement
  - Ran another 30 days to discover…
  - 80 percent of the people in the cohort that was being served an extra ad had started using search engines other than Google as their primary search engine.
- When you see that 80% number, call Twyman's law
- Was it zero for a whole month and jumped to 80%
- What is the OEC here? More ads obviously generate more revenue, but what about negative user impact?

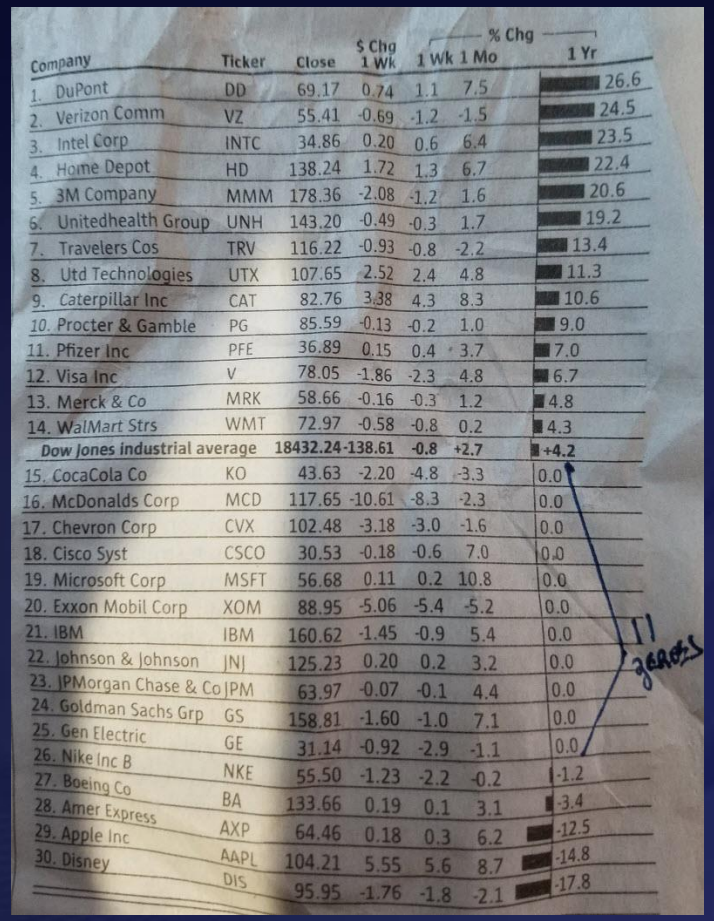
KDD Cup 2000

- **KDD CUP 2000 (article)**
- **Customers who were willing to receive e-mail correlated with heavy spenders (target variable)**
  - Default for registration question was changed from "yes" to "no" on 2/28
  - When it was realized that few were opting-in, the default was changed
  - This coincided with a $10 discount off every purchase
  - Lots of participants found this spurious correlation, but it was terrible for predictions on the test set
- **Sites go through phases (launches) and multiple things change together**



Seattle Times: 0.0% 1-year Change

- On 7/31/2016, Seattle Times published a table of Dow Jones stocks
- 11 out of 30 stocks had 0.0% 1-year change (photo left, tweet)
- Three days later, a correction appeared (photo right)



Company	Ticker	Close	$ Chg 1 Wk	% Chg 1 Wk	% Chg 1 Mo	% Chg 1 Yr
1. DuPont	DD	69.17	0.74	1.1	7.5	26.6
2. Verizon Comm	VZ	55.41	-0.69	-1.2	-1.5	24.5
3. Intel Corp	INTC	34.86	0.20	0.6	6.4	23.5
4. Home Depot	HD	138.24	1.72	1.3	6.7	22.4
5. 3M Company	MMM	178.36	-2.08	-1.2	1.6	20.6
6. Unitedhealth Group	UNH	143.20	-0.49	-0.3	1.7	19.2
7. Travelers Cos	TRV	116.22	-0.93	-0.8	-2.2	13.4
8. Utd Technologies	UTX	107.65	2.52	2.4	4.8	11.3
9. Caterpillar Inc	CAT	82.76	3.38	4.3	8.3	10.6
10. Procter & Gamble	PG	85.59	-0.13	-0.2	1.0	9.0
11. Pfizer Inc	PFE	36.89	0.15	0.4	3.7	7.0
12. Visa Inc	V	78.05	-1.86	-2.3	4.8	6.7
13. Merck & Co	MRK	58.66	-0.16	-0.3	1.2	4.8
14. WalMart Strs	WMT	72.97	-0.58	-0.8	0.2	4.3
Dow Jones industrial average		18432.24	-138.61	-0.8	+2.7	+4.2
15. CocaCola Co	KO	43.63	-2.20	-4.8	-3.3	0.0
16. McDonalds Corp	MCD	117.65	-10.61	-8.3	-2.3	0.0
17. Chevron Corp	CVX	102.48	-3.18	-3.0	-1.6	0.0
18. Cisco Syst	CSCO	30.53	-0.18	-0.6	7.0	0.0
19. Microsoft Corp	MSFT	56.68	0.11	0.2	10.8	0.0
20. Exxon Mobil Corp	XOM	88.95	-5.06	-5.4	-5.2	0.0
21. IBM	IBM	160.62	-1.45	-0.9	5.4	0.0
22. Johnson & Johnson	JNJ	125.23	0.20	0.2	3.2	0.0
23. JPMorgan Chase & Co	JPM	63.97	-0.07	-0.1	4.4	0.0
24. Goldman Sachs Grp	GS	158.81	-1.60	-1.0	7.1	0.0
25. Gen Electric	GE	31.14	-0.92	-2.9	-1.1	0.0
26. Nike Inc B	NKE	55.50	-1.23	-2.2	-0.2	-1.2
27. Boeing Co	BA	133.66	0.19	0.1	3.1	-3.4
28. Amer Express	AXP	64.46	0.18	0.3	6.2	-12.5
29. Apple Inc	AAPL	104.21	5.55	5.6	8.7	-14.8
30. Disney	DIS	95.95	-1.76	-1.8	-2.1	-17.8

Dow30 stocks-table correction

A table showing the stock performance of the 30 companies in the Dow Jones industrial average, which ran in the Sunday Business section, included incorrect numbers for the one-year percentage change. Here is the correct data as of Friday's close.

Company	Ticker	Close	$ Chg 1 Wk	% Chg 1 Wk	% Chg 1 Mo	% Chg 1 Yr
1. Johnson & Johnson	JNJ	125.23	0.20	0.2	3.2	28.0
2. DuPont	DD	69.17	0.74	1.1	7.5	26.8
3. Microsoft Corp	MSFT	56.68	0.11	0.2	10.8	24.4
4. Intel Corp	INTC	34.86	0.20	0.6	6.4	23.9
5. Verizon Comm	VZ	55.41	-0.69	-1.2	-1.5	23.3
6. Gen Electric	GE	31.14	-0.92	-2.9	-1.1	22.8
7. McDonalds Corp	MCD	117.65	-10.61	-8.3	-2.3	21.3
8. 3M Company	MMM	178.36	-2.08	-1.2	1.6	20.7
9. Chevron Corp	CVX	102.48	-3.18	-3.0	-1.6	20.7
10. Home Depot	HD	138.24	1.72	1.3	6.7	20.3
11. Unitedhealth Group	UNH	143.20	-0.49	-0.3	1.7	19.7
12. Exxon Mobil Corp	XOM	88.95	-5.06	-5.4	-5.2	16.0
13. Procter & Gamble	PG	85.59	-0.13	-0.2	1.0	15.1
14. Travelers Cos	TRV	116.22	-0.93	-0.8	-2.2	11.9
15. Cisco Syst	CSCO	30.53	-0.18	-0.6	7.0	10.7
16. Utd Technologies	UTX	107.65	2.52	2.4	4.8	9.9
17. CocaCola Co	KO	43.63	-2.20	-4.8	-3.3	9.5
18. Caterpillar Inc	CAT	82.76	3.38	4.3	8.3	9.2
19. Pfizer Inc	PFE	36.89	0.15	0.4	3.7	5.5
20. Visa Inc	V	78.05	-1.86	-2.3	4.8	4.3
Dow Jones industrial average		18432.24	-138.61	-0.8	+2.7	+4.2
21. WalMart Strs	WMT	72.97	-0.58	-0.8	0.2	4.1
22. Merck & Co	MRK	58.66	-0.16	-0.3	1.2	2.6
23. IBM	IBM	160.62	-1.45	-0.9	5.4	2.4
24. Nike Inc B	NKE	55.50	-1.23	-2.2	-0.2	-2.6
25. JPMorgan Chase & Co	JPM	63.97	-0.07	-0.1	4.4	-4.0
26. Boeing Co	BA	133.66	0.19	0.1	3.1	-4.5
27. Apple Inc	AAPL	104.21	5.55	5.6	8.7	-12.3
28. Amer Express	AXP	64.46	0.18	0.3	6.2	-13.7
29. Disney	DIS	95.95	-1.76	-1.8	-2.1	-18.9
30. Goldman Sachs Grp	GS	158.81	-1.60	-1.0	7.1	-21.3

Alaska Airlines Flights from Adak Island

- People searching flights out of Adak Island suddenly?
- No, never set default to first alphabetical entry

- https://twitter.com/ronnyk/status/524600225093009408

Applying Bayes Rule to Breakthrough Results

- Bing's north-star metric is Sessions/user, but improving it in controlled experiments is extremely rare
- Let's assume that the distribution we see in experiments is Normal, centered on 0, with a standard-deviation of 0.25%
- If an experiment shows +2.0% improvement to Sessions/user, we will call out Twyman's law, pointing out that 2.0% is "extremely interesting" but also eight standard-deviations from the mean, and thus has a probability of 1e-15 excluding other factors.
- Even with a statistically significant result, the prior is so strong against this result, that we avoid any celebration and start working on finding the bug, which is usually an instrumentation error
- Replication is key in such cases.
  See Seven Rules of Thumb for Web Site Experimenters, rule #2

P=NP Proofs

- Twyman's law is regularly applied to proofs that $P = NP$
- No modern editor will celebrate such a submission against a strong prior that a correct proof is very unlikely
- Instead, they will send it to a reviewer to find the bug, attaching a template that says "with regards to your proof that $P = NP$, the first major error is on page x."
- See Seven Rules of Thumb for Web Site Experimenters, rule #2