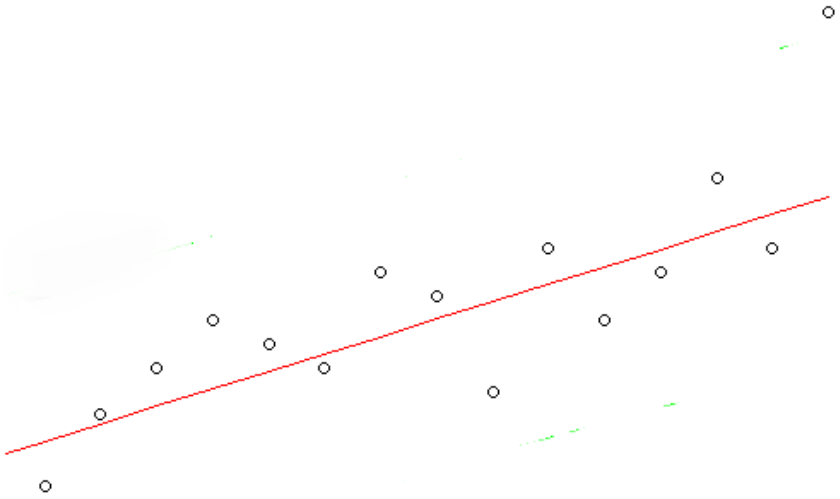
Jason Crease Follow
Jan 2 · 6 min read

Was 2016 especially dangerous for celebrities? An empirical analysis.



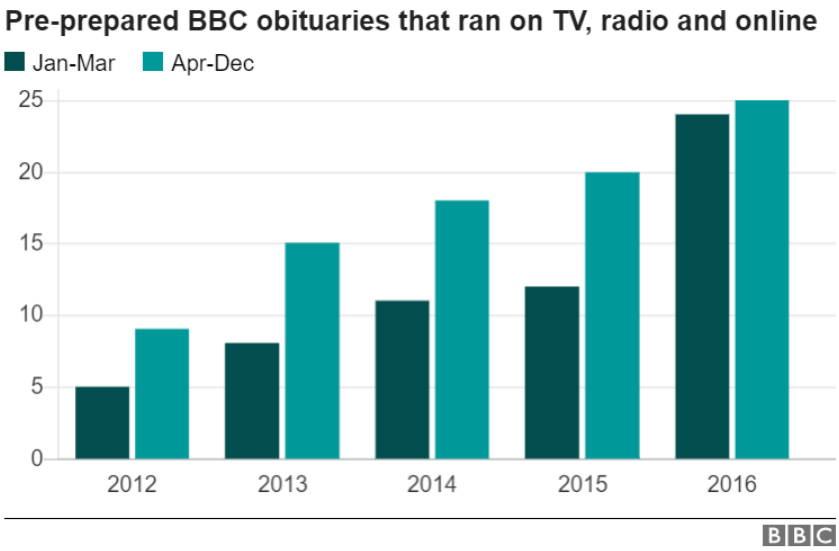
This will explained later. Guess which one is 2016?

It's become cliché that unusually many prominent people died in 2016. Is this true? To answer this we need to know:

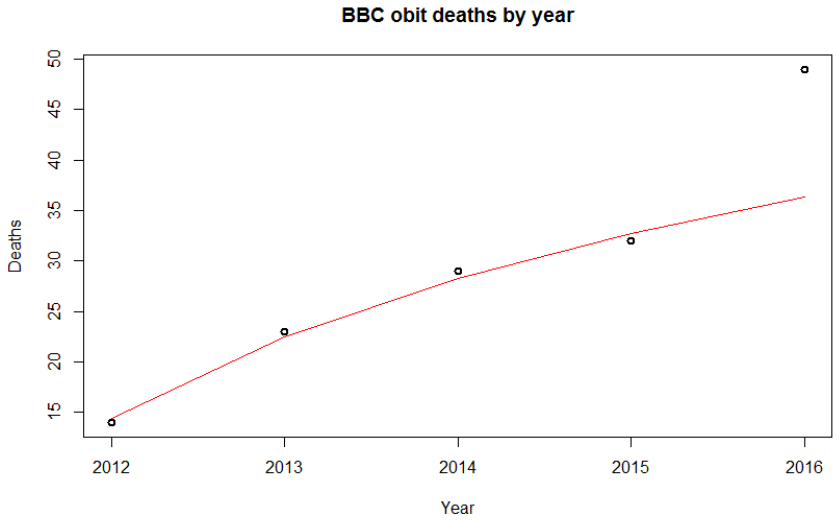
1. *(The easy part)* What is ***unusually many***?
2. *(The hard part)* What is a ***celebrity***?

The BBC analysis

For their analysis, the BBC defined celebrities as those with a pre-prepared obituary. That is, a pre-written ready-to-run obituary. Given this definition, it certainly looks like an usually high number of prominent people died in 2016:



But couldn't this just be due to an increasing number of pre-prepared obits, or some other long-term trend? You can try to account for this by interpolating from 2012 to 2015 (I used a logarithmic trend— a quadratic gave similar results). Thus, I'd expect **36.4** celebrities to die in 2016. **49** did.



Using the obvious Poisson interpretation, P(Deaths ≥ 49) = **0.026**. So a **1 in 40 year freakiness**.

Just taking January to April gives an even more extreme picture. I'd predict **13.7** deaths—instead there were **24**. This has a probability of

just **0.007**. The specific choice of January to April stinks of data-dredging, but I'm still kinda impressed.

Wikipedia and prominence

I'm unsatisfied with the pre-prepared BBC obit as a metric of celebrity:

1. It has a British bias (although it's obviously impossible to be entirely objective.)
2. When do they prepare obits? Maybe they just happened to write a load in December 2015.
3. The decision to prepare an obit still remains the subjective opinion of a few bods at the BBC.
4. Maybe the 2016 deaths were merely unusually *expected*, thus had obits ready.

Wikipedia to the rescue!

Maybe Wikipedia biographies would be a good source? Noteworthy people should have long and carefully-tended articles.

My analysis is similar to the book Who's Bigger?. You may just want to skip my article and read that book.

Using C#, the Wikipedia API, and plenty of regexes, I extracted a list of prominent deaths from each year's summary page, eg https://en.wikipedia.org/wiki/1992#Deaths . This gives a total of 6475 people, or roughly 20 a month. Then I used the Wikipedia API to get the lengths of these biographies in bytes, and the number of revisions per article.

I probably hit the web API pretty hard, so I made a small donation out of guilt :(.

Article length and revisions as a measure of prominence

For those dying since 1987, these are the 11 longest biographies. Note I'm only using the English Wikipedia:

Name	Year	Length
Johan Cruyff	2016	273366
Pope John Paul II	2005	261191
Bobby Fischer	2008	233183
Michael Jackson	2009	231669
Ronald Reagan	2004	224664
Hugo Chávez	2013	218817
Frank Sinatra	1998	206283
Whitney Houston	2012	204658
Muhammad Ali	2016	195950
Ted Kennedy	2009	194944
Nelson Mandela	2013	190788

This is kinda unsatisfactory. Johan Cruyff's long football career gives him a long, detailed article, but is he really more significant than Michael Jackson? Michael Jackson has 8x as many revisions as Johan Cruyff, I presume this is because people pay him 8x as much attention.

These are the 20 articles with the most revisions:

Name	Year	Edits
Michael Jackson	2009	28685
Ronald Reagan	2004	17864
Pope John Paul II	2005	16438
Diana, Princess of Wales	1997	14345
Fidel Castro	2016	13937
Osama bin Laden	2011	13768
Freddie Mercury	1991	13277
Hugo Chávez	2013	13232
Saddam Hussein	2006	12377
Tupac Shakur	1996	11841
David Bowie	2016	11581
Whitney Houston	2012	11410
Kurt Cobain	1994	11258
Johnny Cash	2003	11192
George Harrison	2001	10799
Muhammad Ali	2016	10795
Richard Nixon	1994	10692
Margaret Thatcher	2013	10616
Steve Jobs	2011	10588
Frank Sinatra	1998	10485

Ah, that's better! Every one a mega-celebrity. Note three are from 2016.

But now I found is a bias towards contentious figures (such as Indian guru Sathya Sai Baba), and those whom the *man in the street* has a lot to say about. Some important long-dead figures have good biographies that were rapidly and conclusively written in a few sessions by scholars—surely they deserve recognition?

A few other random bits:

- The longest biography on Wikipedia is of Belgian astronomer Eric Walter Elst. It tediously lists thousands of asteroids that he discovered, but has few revisions.

- When plotting Revisions against Lengths, we can see that there is a good correlation between Revisions and Lengths. The Spearman rank correlation-cofficient is 0.884—quite high.
- Looking at revisions and lengths there is an exponential trend. That is, something like 80% of the length/revisions is in 20% of the articles.
- Most Wikipedia editors are American, male, nerdy, and young. I suspect.
- I'm only using the English Wikipedia. My analysis is Anglocentric. And US-centric.

My definition of celebrity

Neither article-length nor number-of-revisions seems ideal. Therefore I define one's ***Celebrity*** as the harmonic mean of the logarithms of your article-length and number-of-revisions, each normalised by the maximum you can achieve in each category.

$$Celebrity = \frac{2}{\frac{log(273366)}{log(Length)} + \frac{log(28685)}{log(Revisions)}}$$

A maximal celebrity will score 1.0. Unknowns will score 0.0.

The harmonic average has the nice property that it biases against those with unusually high scores for *Length* or *Revisions*. So a person with a very long article that has only been revised a few times is probably an anomaly, and will score poorly. Likewise, a short biography that has been heavily revised will also score poorly.

Here are my top-30 based on this metric:

	Name	Year	Length	Edits	Celebrity		Name	Year	Length	Edits	Celebrity
1	Michael Jackson	2009	231669	28685	0.993	***16***	Prince (musician)	2016	153721	10067	0.925
2	Pope John Paul II	2005	261191	16438	0.970	***17***	Freddie Mercury	1991	107243	13277	0.925
3	Ronald Reagan	2004	224664	17864	0.969	***18***	George Harrison	2001	135082	10799	0.924
4	Hugo Chávez	2013	218817	13232	0.953	***19***	Tupac Shakur	1996	119664	11841	0.924
5	Fidel Castro	2016	169282	13937	0.945	***20***	Bobby Fischer	2008	233183	7388	0.924
6	Whitney Houston	2012	204658	11410	0.942	***21***	Saddam Hussein	2006	112276	12377	0.923
7	Diana, Princess of Wales	1997	140458	14345	0.940	***22***	Ted Kennedy	2009	194944	8177	0.923
8	Frank Sinatra	1998	206283	10485	0.938	***23***	Aaliyah	2001	142535	10128	0.923
9	Muhammad Ali	2016	195950	10795	0.938	***24***	Ted Bundy	1989	138701	9937	0.921
10	Steve Jobs	2011	186686	10588	0.935	***25***	Gerald Ford	2006	155864	9043	0.920
11	Osama bin Laden	2011	130728	13768	0.935	***26***	Muammar Gaddafi	2011	151728	9127	0.920
12	David Bowie	2016	163149	11581	0.935	***27***	Stanley Kubrick	1999	141464	8837	0.915
13	Nelson Mandela	2013	190788	9941	0.933	***28***	Amy Winehouse	2011	157377	7824	0.913
14	Margaret Thatcher	2013	170506	10616	0.932	***29***	Johnny Cash	2003	93352	11192	0.911
15	Richard Nixon	1994	154879	10692	0.929	***30***	Kurt Cobain	1994	85445	11258	0.908

These seems like a nice compromise between the two metrics.

Here's a few other random mega-celebrities for comparison:

Name	Length	Edits	Celeb
Jesus	204333	29165	0.989
Adolf Hitler	160758	25127	0.972
Elvis Presley	201974	17954	0.965
Albert Einstein	138015	16023	0.944
Justin Bieber	110973	8029	0.901

I now make two convenient definitions: a ***P200*** and a ***P1000,*** a mega-celebrity and celebrity respectively. Note that every P200 is also a P1000.

- You're a ***P200*** if you're in the top 200 of my list, for those dying 2000–2016. Just making it into P200 territory are Enoch Powell and Edward Heath.
- You're a ***P1000*** (or ***P1K)*** if you're in top 1000. Just making it are Dom DeLuise and Jeff Hanneman.

Prominent People's deaths in 2016 on Wikipedia

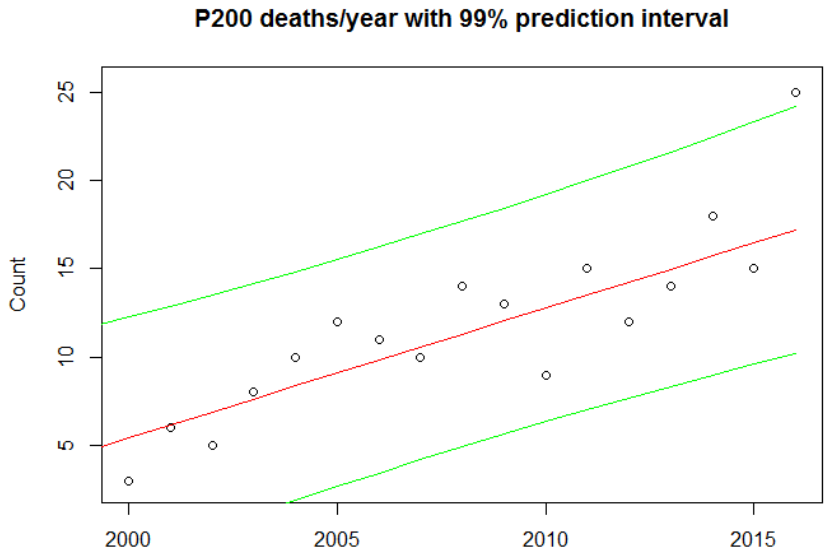
All right, time to look at 2016.

Looking at P200 and P1Ks there appears to be a long-term linear trend. I guess this is because articles of living celebrities are continuously expanded, so recently-dead celebrities have longer articles.

Indeed, Wikipedia statistics show linear increases in most metrics since 2010. I think it's reasonable to do a linear interpolation of 2000–2015, and use this to predict 2016.

P200s

I would predict **17** P200 deaths in 2016. There were actually **25**. This is just outside the 99.5% prediction interval. So roughly **a once-in-200-years event**.

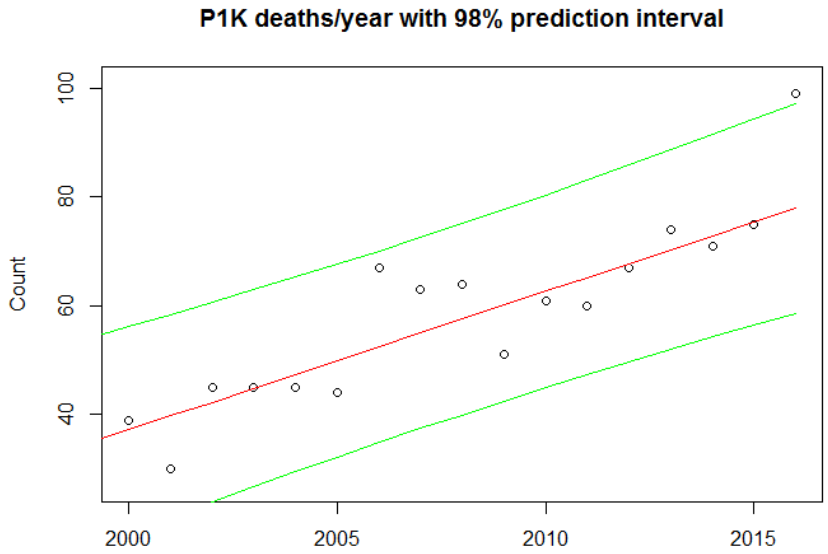


2016's P200s were: *Fidel Castro, Muhammad Ali, David Bowie, Prince, George Michael, Johan Cruyff, Bhumibol Adulyadej, Leonard Cohen, Antonin Scalia, Elie Wiesel, Nancy Reagan, John Glenn, Carrie Fisher, Chyna, Harper Lee, Kimbo Slice, Ernst Nolte, Rob Ford, Pierre Boulez, Alan Rickman, Shimon Peres, Christina Grimmie, Terry Wogan, Abbas Kiarostami,* and *Merle Haggard.*

(Do you also feel like uncultured scum for not knowing who Abbas Kiarostami was? I think I'm glad I'd never heard of Chyna though.)

P1Ks

I predict **78** P1K deaths in 2016. There were actually **99**, which is roughly at the 99% prediction interval. So again roughly a **once-in-a-century event**.



Technical note: Deaths are Poisson distributed, not normal! What's with this linear-least-squares rubbish?

It looks like a reasonable assumption that the Poisson parameter (deaths-per-year) increases linearly with time:

$\lambda(t) = at + b$

Due to the central-limit theorem, the sample-mean (i.e. observed deaths per year) of a Poisson approaches a Gaussian. So doing linear-least-squares regression assuming Gaussian-residuals on the Poisson-parameter/observed-deaths variable could be fine in-the-limit.

However, since λ itself increases with time, the residuals will increase in magnitude. Additionally, the normal-approximation is poor for small λ, especially at the tails, which is where we are.

So let's redo the maths with the Poisson CDFs. Taking the λs from the earlier linear-interpolation:

For P200s: $P(D \geq 25 \mid \lambda = 17) = 0.04$

For P1000s: $P(D \geq 99 \mid \lambda = 78) = 0.01$

It still looks like an unusually high number of celebrities died, but the number of mega-celebrity deaths was less surprising that the large number of rank-and-file celebrities.

Conclusion

2016 was indeed a year of surprisingly-many celebrity deaths.