

Real life examples of distributions with negative skewness

Asked 5 years, 8 months ago Active 1 year, 3 months ago Viewed 125k times



20



13

Inspired by "[real-life examples of common distributions](#)", I wonder what pedagogical examples people use to demonstrate negative skewness? There are many "canonical" examples of symmetric or normal distributions used in teaching - even if ones like height and weight don't survive closer biological scrutiny! Blood pressure might be nearer normality. I like astronomical measurement errors - of historic interest, they are intuitively no more likely to lie in one direction than another, with small errors more likely than large.

Common pedagogical examples for positive skewness include people's incomes; mileage on used cars for sale; reaction times in a psychology experiment; house prices; number of accident claims by an insurance customer; number of children in a family. Their physical reasonableness often stems from being bounded below (usually by zero), with low values being plausible, even common, yet very large (sometimes orders of magnitude higher) values are well-known to occur.

For negative skew, I find it harder to give unambiguous and vivid examples that a younger audience (high schoolers) can intuitively grasp, perhaps because fewer real-life distributions have a clear upper bound. A bad-taste example I was taught at school was "number of fingers". Most folk have ten, but some lose one or more in accidents. The upshot was "99% of people have a higher-than-average number of fingers"! [Polydactyly](#) complicates the issue, as ten is not a strict upper bound; since both missing and extra fingers are rare events, it may be unclear to students which effect predominates.

I usually use a binomial distribution with high p . But students often find "number of satisfactory components in a batch is negatively skewed" less intuitive than the complementary fact that "number of faulty components in a batch is positively skewed". (The textbook is industrially themed; I prefer cracked and intact eggs in a box of twelve.) Maybe students feel that "success" should be rare.

Another option is to point out that if X is positively skewed then $-X$ is negatively skewed, but to place this in a practical context ("negative house prices are negatively skewed") seems doomed to pedagogical failure. While there are benefits to teaching the effects of data transformations, it seems wise to give a concrete example first. I would prefer one that does not seem artificial, where the negative skew is quite unambiguous, and for which students' life-experience should give them an awareness of the shape of the distribution.

distributions

skewness

teaching

-
- 4 It is not apparent that negating a variable will be a "pedagogical failure," because there is the option of adding a constant without changing the shape of the distribution. Many skewed distributions involve proportions X for instance, and the complementary proportions $1 - X$ are usually just as natural and easy to interpret as the original proportions. Even with house prices X the values $C - X$ where C is a maximum house price in the area could be of interest and is not difficult to understand. Also consider using logs and negative power transformations to create negative skew. – [whuber](#) ♦ Mar 7 '14 at 17:30
-
- 2 I agree that $C - X$ in the case of house prices would be a little contrived. But $1/X$ would not: it would be "amount of house you can buy per dollar." I suspect that in any reasonably homogeneous area this would have a strong negative skew. Such examples could teach the deeper lesson that skewness is a function of how we express the data. – [whuber](#) ♦ Mar 7 '14 at 20:13
-
- 3 @whuber It wouldn't be contrived at all. Maximum and minimum *potential* prices in a market arise naturally as those reflecting different evaluations by market participants. Among the buyers, there is conceivably one that would pay maximum price for a given house. And among the sellers there is one that would conceivably accept minimum price. But this information is not public and so actual observed transaction prices are affected by the existence of incomplete information. (CONT'D) – [Alec Papadopoulos](#) Mar 8 '14 at 14:04
-
- 1 CONT'D ... The following paper by Kumbhakar and Parmeter (2010) models exactly that (permitting also the case of symmetry), and with an application on the house market: link.springer.com/article/10.1007/s00181-009-0292-8#page-1 – [Alec Papadopoulos](#) Mar 8 '14 at 14:04
-
- 3 Age at death is negatively skewed in developed countries. – [Nick Cox](#) Mar 12 '14 at 0:59
-

12 Answers



3



In the UK, price of a book. There is a "Recommended retail price" which will generally be the modal price, and virtually nowhere would you have to pay more. But some shops will discount, and a few will discount heavily.

Also, age at retirement. Most people retire at 65-68 which is when the state pension kicks in, very few people work longer, but some people retire in their 50s and quite a lot in their early 60s.

Then too, the number of GCSEs people get. Most kids are entered for 8-10 and so get 8-10. A small number do more. Some of the kids don't pass all their exams though, so there is a steady increase from 0 to 7.

answered Feb 9 '17 at 10:22

community wiki
[user148573](#)



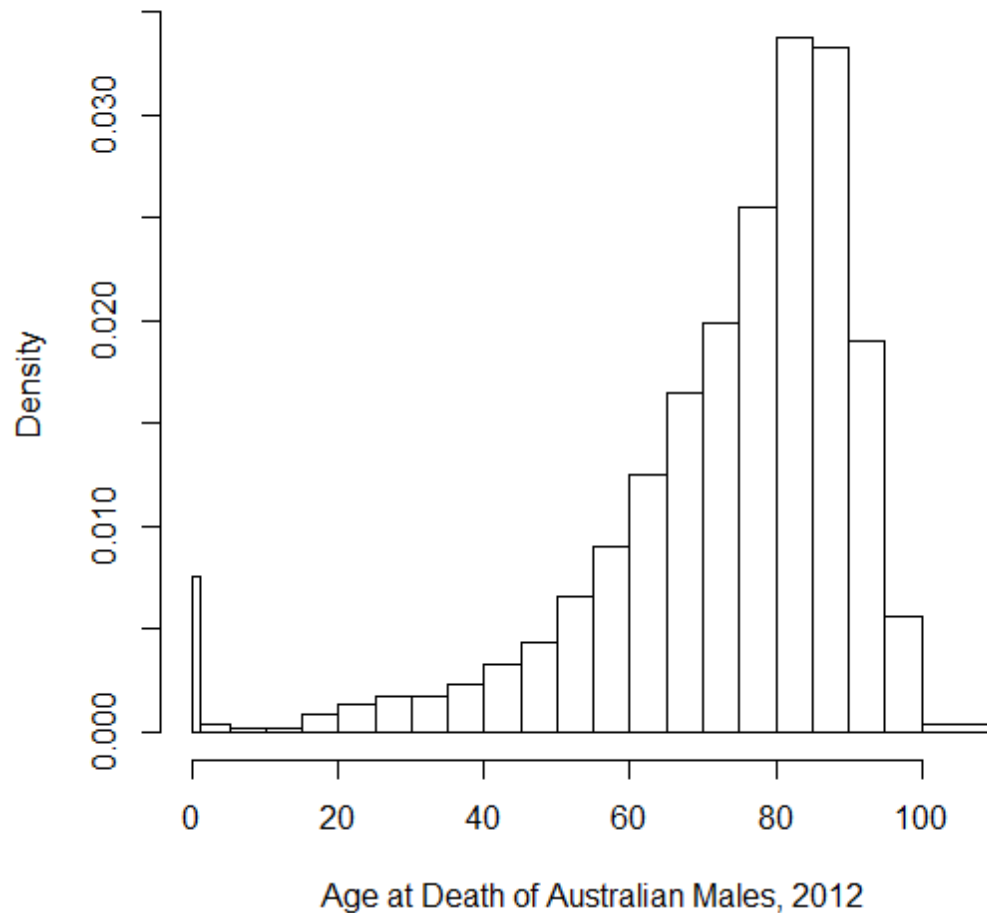
18



Nick Cox accurately commented that "age at death is negatively skewed in developed countries" which I thought was a great example.

I found [the most convenient figures I could lay my hands on](#) came from the Australian Bureau of Statistics ([in particular, I used this Excel sheet](#)), since their age bins went up to 100 year olds and the [oldest Australian male was 111](#) , so I felt comfortable cutting off the final bin at 110 years. Other national statistical agencies often seemed to stop at 95 which made the final bin uncomfortably wide. The resulting histogram shows a very clear negative skew, as well as some other interesting features such as a small peak in death rate among young children, which would be well suited to class discussion and interpretation.

Histogram of Age at Death of Australian Males, 2012



R code with raw data follows, the [HistogramTools package](#) proved very useful for plotting based on aggregated data! Thanks to [this StackOverflow question](#) for flagging it up.

```
library(HistogramTools)
```

```
deathCounts <- c(565, 116, 69, 78, 319, 501, 633, 655, 848, 1226, 1633, 2459,  
3375, 4669, 6152, 7436, 9526, 12619, 12455, 7113, 2104, 241)  
ageBreaks <- c(0, 1, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75,
```

```
breaks = ageBreaks,  
counts = deathCounts,  
xname = "Age at Death of Australian Males, 2012")  
plot(myhist)
```

edited May 23 '17 at 12:39

community wiki

2 revs

Silverfish

-
- 2 Somewhat related to this post, I have heard that retirement ages have negative skewness: most people retire around the nominal age (say, 65 or 67 in many countries) but some (say, workers in coal mines) retire much earlier. – [Christoph Hanck](#) Feb 25 '15 at 5:29
-

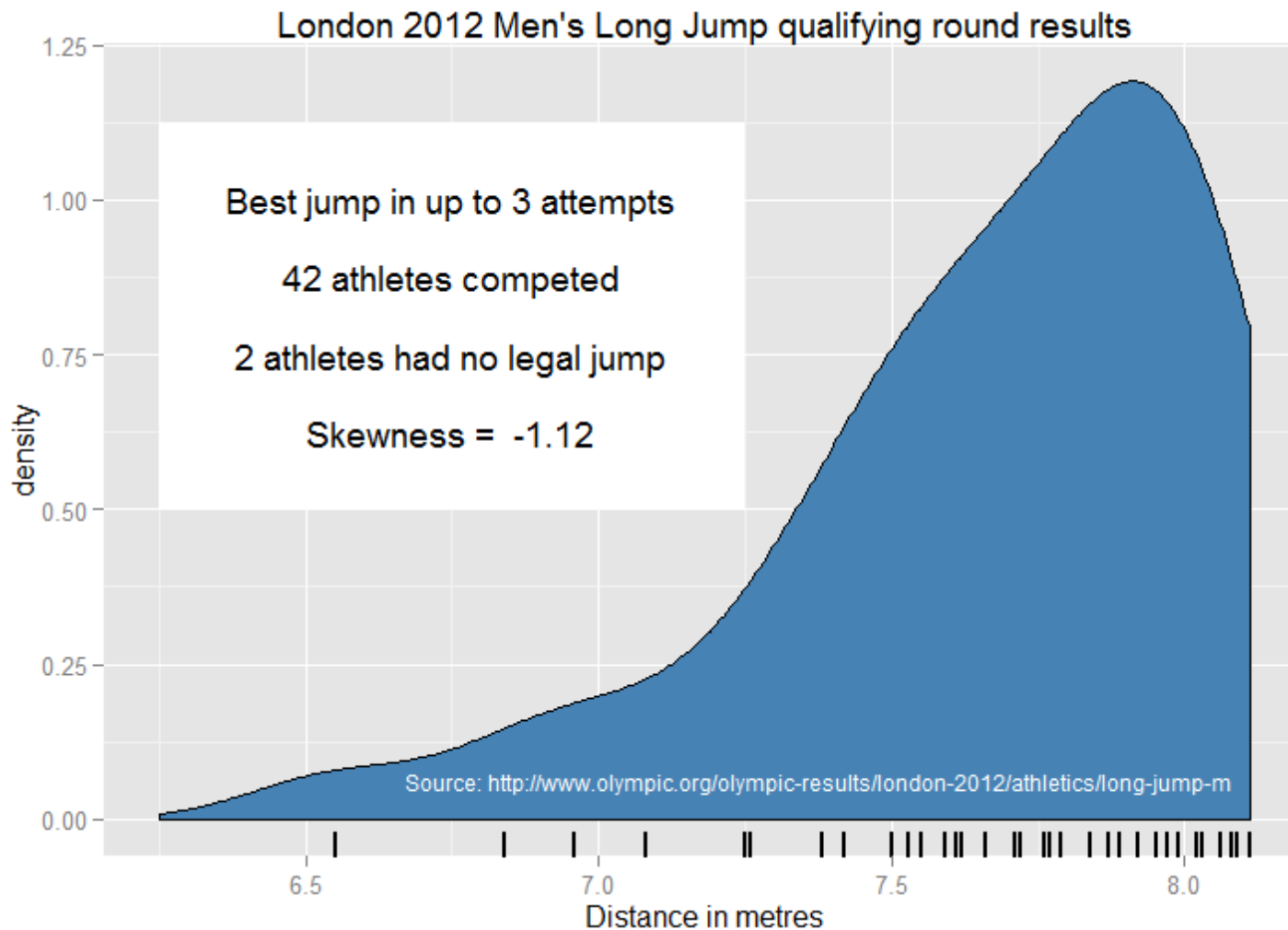
Does the age at death follow some known distribution empirically? – [StubbornAtom](#) Apr 22 '18 at 9:42



11



Here are the results for the forty athletes who successfully completed a legal jump in the qualifying round of the 2012 Olympic men's long jump, presented in a kernel density plot with rug plot underneath.



It seems to be much easier to be a metre behind the main group of competitors than to be a metre ahead, which would explain the negative skewness.

I suspect some of the bunching at the top end is due to the athletes targeting qualification (which required a top twelve finish or a result of 8.10 metres or above) rather than achieving the longest possible distance. The fact that the top two results were 8.11 metres, just above the automatic qualifying mark, is strongly suggestive, as is the way the medal-winning jumps in the Final were both longer and more spread out at 8.31, 8.16 and 8.12 metres. Results in the Final had a slight, non-significant, negative skew.

replicated in the throwing events (shot and javelin) even though they are also events in which a higher number corresponds to a better result. The final points scores were also somewhat negatively skewed.

Data and code

```
require(moments)
require(ggplot2)

sourceAddress <- "http://www.olympic.org/olympic-results/london-
2012/athletics/long-jump-m"

longjump.df <- read.csv(header=TRUE, sep="," , text="
rank,name,country,distance
1,Mauro Vinicius DA SILVA,BRA,8.11
2,Marquise GOODWIN,USA,8.11
3,Aleksandr MENKOV,RUS,8.09
4,Greg RUTHERFORD,GBR,8.08
5,Christopher TOMLINSON,GBR,8.06
6,Michel TORNEUS,SWE,8.03
7,Godfrey Khotso MOKOENA,RSA,8.02
8,Will CLAYE,USA,7.99
9,Mitchell WATT,AUS,7.99,
10,Tyrone SMITH,BER,7.97,
11,Henry FRAYNE,AUS,7.95,
12,Sebastian BAYER,GER,7.92,
13,Christian REIF,GER,7.92,
14,Eusebio CACERES,ESP,7.92,
15,Aleksandr PETROV,RUS,7.89,
16,Sergey MORGUNOV,RUS,7.87,
17,Mohammad ARZANDEH,IRI,7.84,
18,Ignisious GAISAH,GHA,7.79,
19,Damar FORBES,JAM,7.79,
20,Jinzhe LI,CHN,7.77,
21,Raymond HIGGS,BAH,7.76,
22,Alyn CAMARA,GER,7.72,
23,Salim SDIRI,FRA,7.71,
24,Ndiss Kaba BADJI,SEN,7.66,
25,Arsen SARGSYAN,ARM,7.62,
26,Povilas MYKOLAITIS,LTU,7.61,
27,Stanley GBAGBEKE,NGR,7.59,
28,Marcos CHUVA,POR,7.55,
29,Louis TSATOUMAS,GRE,7.53,
30,Stepan WAGNER,CZE,7.50,
31,Viktor KUZNYETSOV,UKR,7.50,
32,Luis RIVERA,MEX,7.42
```

```

36,Xiaoyi ZHANG,CHN,7.25,
37,Mohamed Fathalla DIFALLAH,EGY,7.08,
38,Roman NOVOTNY,CZE,6.96,
39,George KITCHENS,USA,6.84,
40,Vardan PAHLEVANYAN,ARM,6.55,
NA,Luis MELIZ,ESP,NA,
NA,Irving SALADINO,PAN,NA")

roundedSkew <- signif(skewness(longjump.df$distance, na.rm=TRUE), 3)

ggplot(longjump.df, aes(x=distance)) +
  xlab("Distance in metres") +
  ggtitle("London 2012 Men's Long Jump qualifying round results") +
  geom_rug(size=0.8) +
  geom_density(fill="steelblue") +
  annotate("text", x=7.375, y=0.0625, colour="white", label=paste("Source:",
sourceAddress), size=3) +
  annotate("rect", xmin = 6.25, xmax = 7.25, ymin = 0.5, ymax = 1.125,
fill="white") +
  annotate("text", x=6.75, y=1, colour="black", label="Best jump in up to 3
attempts") +
  annotate("text", x=6.75, y=.875, colour="black", label="42 athletes
competed") +
  annotate("text", x=6.75, y=.75, colour="black", label="2 athletes had no
legal jump") +
  annotate("text", x=6.75, y=.625, colour="black", label=paste("Skewness = ",
roundedSkew))

# Results of the top twelve who qualified for the Final were closer to symmetric
skewness(longjump.df$distance[1:12])
# -0.1248782

# Results in the Final (some had 3 jumps, others 6) were only slightly
negatively skewed
skewness(c(8.31, 8.16, 8.12, 8.11, 8.10, 8.07, 8.01, 7.93, 7.85, 7.80, 7.78,
7.70))
# -0.08578357

# Compare to Seoul 1988 Heptathlon
require(HSAUR)
skewness(heptathlon)

```

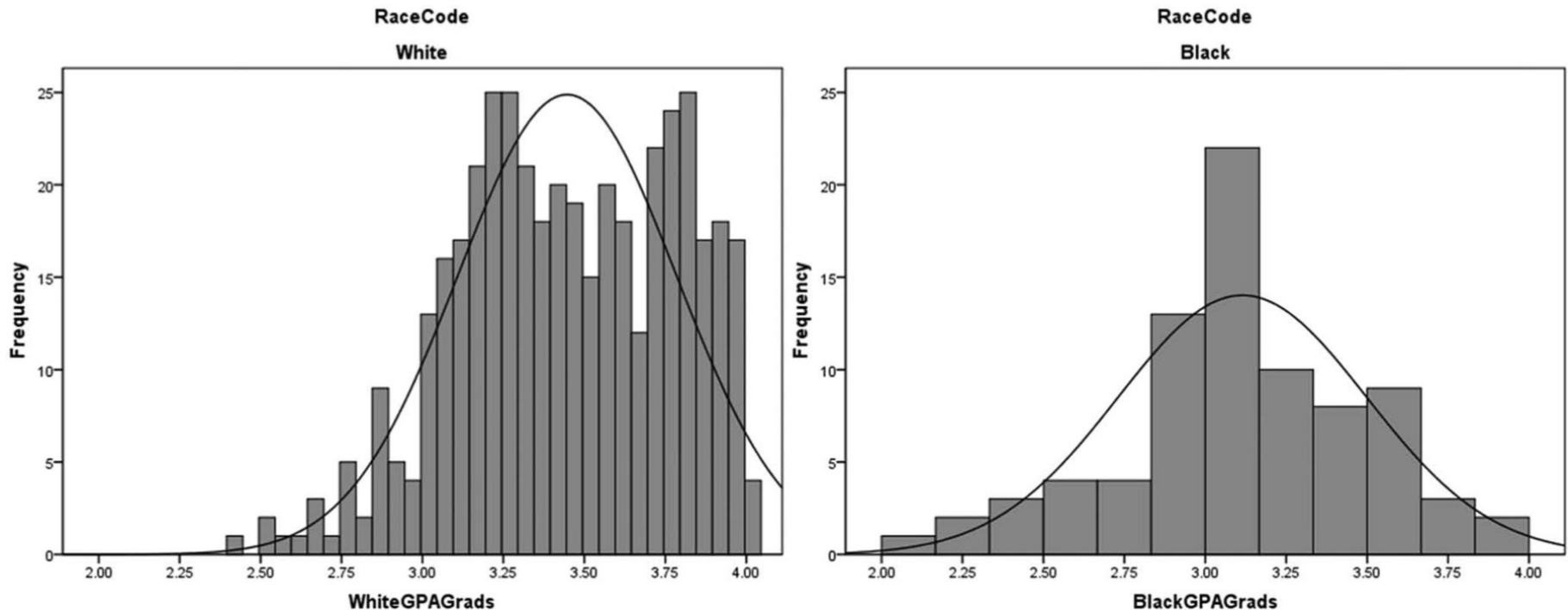
edited Jan 28 '15 at 10:19

community wiki
2 revs

Scores on easy tests, or alternatively, scores on tests for which students are especially motivated, tend to be left skew.

As a result, the SAT/ACT scores of students entering sought after colleges (and even more so, their GPAs) tend to be left skew. There's plenty of examples at collegeapps.about.com [e.g. a plot of University of Chicago SAT/ACT and GPA is here](#).

Similarly GPAs of graduates are often left-skew, e.g. the histograms below of GPAs of white and black graduates at a for-profit university taken from Fig 5 of Gramling, Tim. "[How five student characteristics accurately predict for-profit university graduation odds](#)." *SAGE Open* 3.3 (2013): 2158244013497026.



(It's not hard to find other, similar examples.)

edited Feb 4 '15 at 23:33

community wiki
2 revs, 2 users 79%
Glen_b

- 2 For an introductory stats class I think this example works well pedagogically - it is something students are likely to have real-life experience of, can reason about intuitively, and can confirm against widely available data sets. [Glen_b](#), Mar 11 '14 at 0:47

In Stochastic Frontier Analysis, and specifically in its historically initial focus, production, the production function of a firm/production unit in general, is specified stochastically as

$$q = f(\mathbf{x}) + u - w$$

where q is the actual output produced by the firm, and $f(\mathbf{x})$ is its production function (which is understood more as an input-output relation rather than a mathematical expression reflecting "engineering" relations) with \mathbf{x} being a vector of production inputs (capital, labor, energy, materials, etc). The production function in Economic Theory represents *maximum* output, given technology and inputs, i.e. it embodies *full efficiency*. Then u is a zero-mean normal disturbance on the production process, and w is a non-negative random variable representing *deviation from full efficiency* due to reasons that the econometrician may not know, but he can measure through this set up. This random variable is usually assume to follow a half-normal or exponential distribution. Assuming the half normal (for a reason), we have

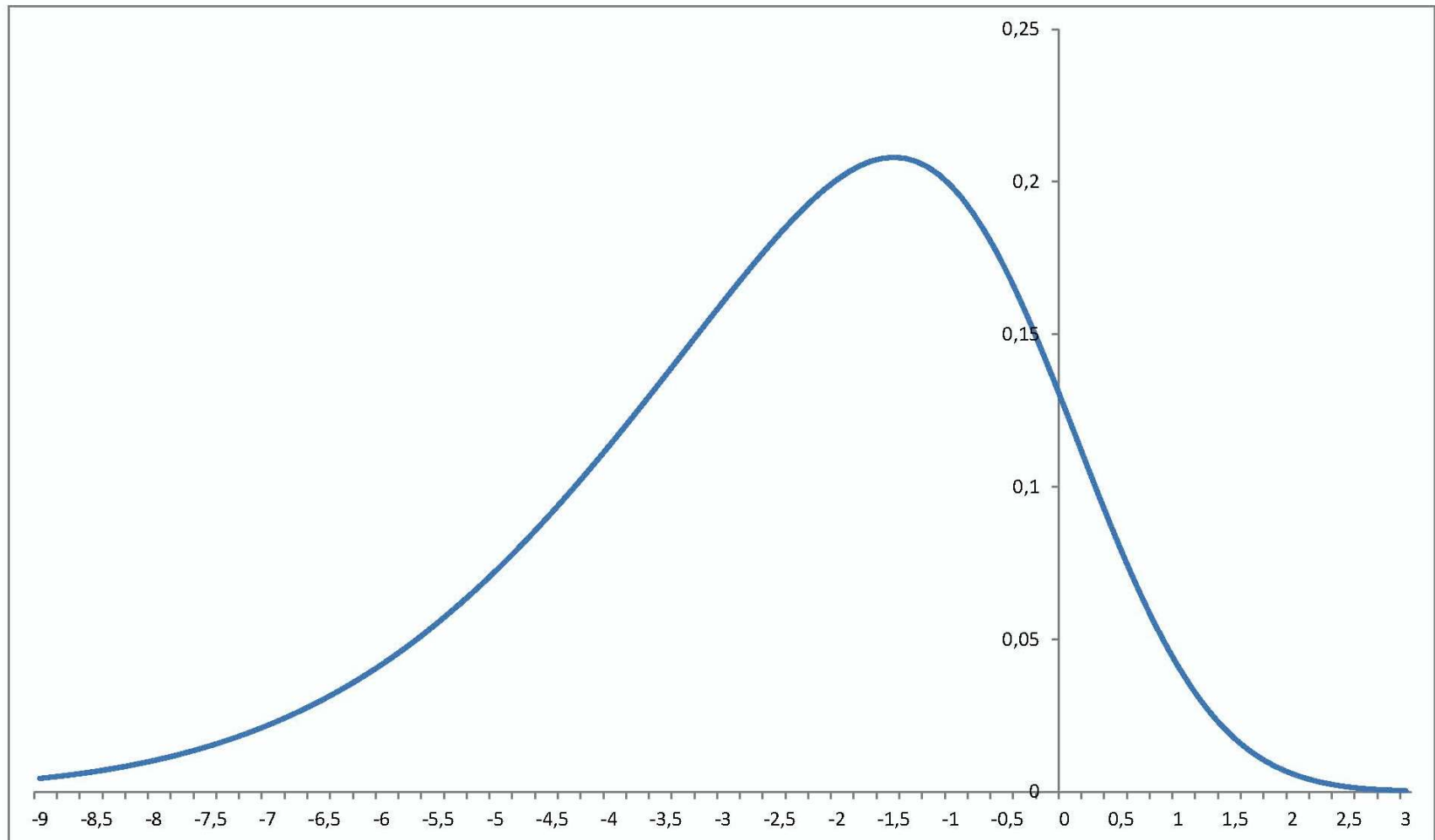
$$u \sim N(0, \sigma_u^2), \quad w \sim HN\left(\sqrt{\frac{2}{\pi}}\sigma_2, \left(1 - \frac{2}{\pi}\right)\sigma_2^2\right)$$

where σ_2 is the standard deviation of the "underlying" normal random variable whose absolute value is the Half-normal.

The composite error-term $\varepsilon = u - w$ is characterized by the following density

$$f_\varepsilon(\varepsilon) = \frac{2}{s_2} \phi(\varepsilon/s_2) \Phi\left(-\frac{\sigma_2}{\sigma_u} \cdot (\varepsilon/s_2)\right), \quad s_2^2 = \sigma_u^2 + \sigma_2^2$$



This is a skew-normal density, with location parameter 0, scale parameter s_2 and skew parameter $(-\frac{\sigma_2}{\sigma_u})$, where ϕ and Φ are the standard normal pdf and cdf respectively. For $\sigma_u = 1$, $\sigma_2 = 3$, the density looks like this:



So negative skewness is, I'd say, the most natural modelling of the efforts of human race itself: always deviating from its imagined ideal - in most cases lagging behind it (the negative part of the density), while in relatively fewer cases, transcending its perceived limits (the positive part of the density). *Students themselves* can be modeled as such a production function. It is straightforward to map the symmetric disturbance and the one-sided error to aspects of real life. I cannot imagine how more intuitive can one get about it.

edited Mar 8 '14 at 11:22

community wiki

-
- 1 This answer seems to echo @Glen_b's suggestion of grad GPA. Highly motivated human behavior aimed at an elusive ideal certainly fits that scenario! Efficiency in general is a great example. – [Nick Stauner](#) Mar 8 '14 at 10:03 
 - 2 @Nick Stauner The important point here is that we consider "actual minus target" signed, not the "distance" in absolute values. We keep the sign in order to know whether we are above or below the target. The intuition here is, exactly as you write, that "highly motivated" behavior will push "actual" closer to "target", creating asymmetry. – [Alec Papadopoulos](#) Mar 8 '14 at 11:12 
 - 1 @NickStauner Indeed, Silverfish's own post of long jump qualifying results also relates to 'highly motivated behavior' (considering limits of what humans can presently achieve as a kind of informal 'elusive ideal') – [Glen_b](#) Jan 28 '15 at 2:42
-



6

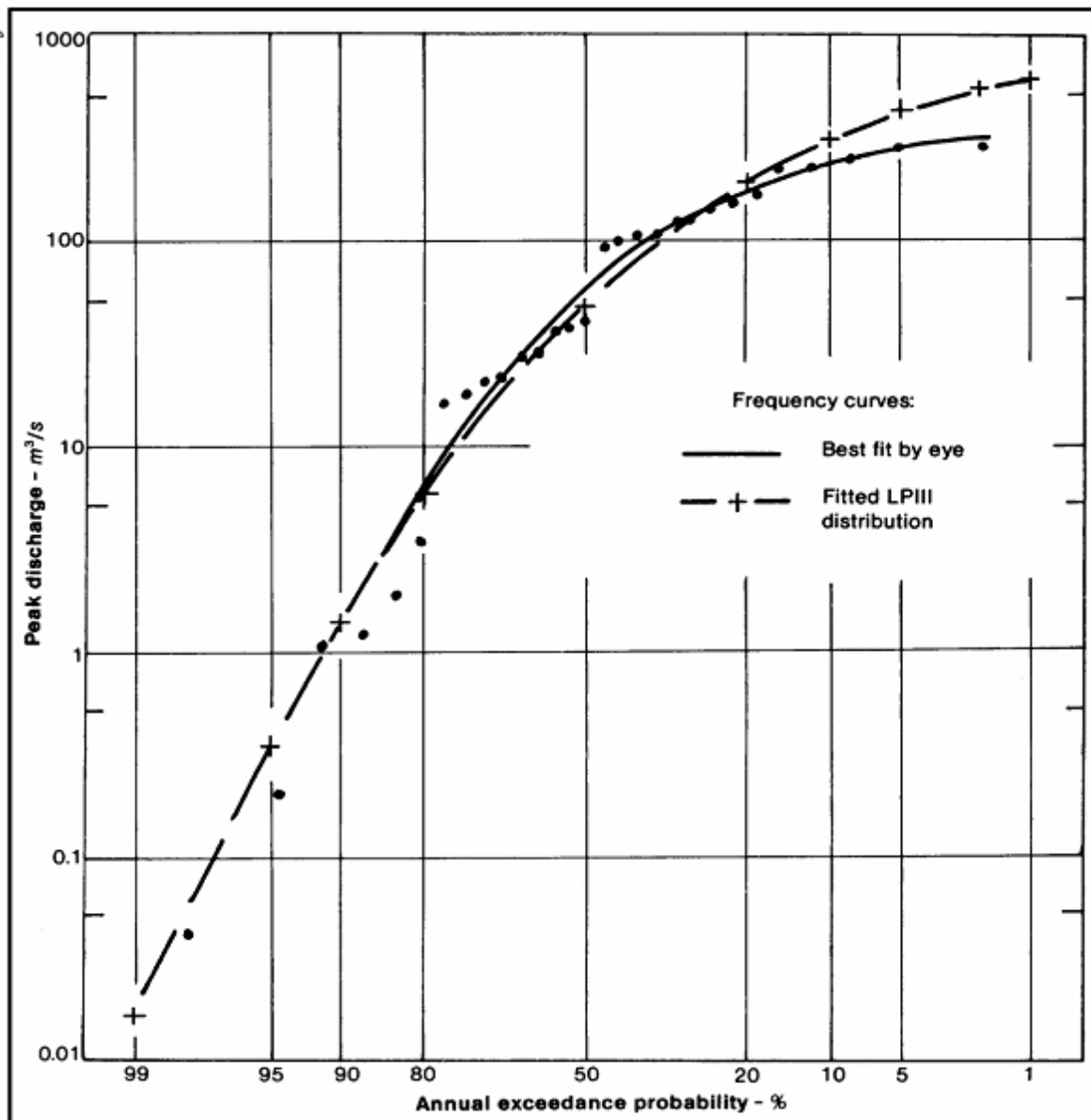


Negative skewness is common in flood hydrology. Below is an example of a flood frequency curve (South Creek at Mulgoa Rd, lat -33.8783, lon 150.7683) which I've taken from 'Australian Rainfall and Runoff' (ARR) the guide to flood estimation developed by Engineers, Australia.

There is a comment in ARR:

With negative skew, which is common with logarithmic values of floods in Australia, the log Pearson III distribution has an upper bound. This gives an upper limit to floods that can be drawn from the distribution. In some cases this can cause problems in estimating floods of low AEP, but often causes no problems in practice. [Extracted from Australian Rainfall and Runoff - Volume 1, Book IV Section 2.]

Often floods, at a particular location, are considered to have an upper bound called the 'Probable Maximum Flood' (PMF). There are standard ways of calculating a PMF.



answered Mar 11 '14 at 22:39

community wiki
Tony Ladson

its distribution negatively (merely by taking a suitably negative Box-Cox parameter). It all comes down to what is meant by "easily grasped," I suppose-- but that's a question about the students, not about statistics. – [whuber](#) ♦ Mar 11 '14 at 23:08

5

Asset price changes (returns) typically have negative skew - many small price increases with a few large price drops. The skew seems to hold for almost all types of assets: stocks prices, commodity prices, etc. The negative skew can be observed in monthly price changes but is much more evident when you start looking at daily or hourly price changes. I think this would be a good example because you can show the effects of frequency on skew.

More details: <http://www.fusioninvesting.com/2010/09/what-is-skew-and-why-is-it-important/>

answered Mar 7 '14 at 18:40

community wiki
[wcampbell](#)

I like this example a lot! Is there an intuitive way of explaining it - essentially, "downside shocks are more likely (or at least, likely to be more severe) than upside shocks"? – [Silverfish](#) Mar 7 '14 at 19:17

-
- 2 @Silverfish I would phrase it as extreme negative market outcomes are more likely than extreme positive market outcomes. Markets also have asymmetric volatility. Market volatility generally increases more following negative returns than positive returns. This is often modeled with Garch models, such as GJR-Garch (see Arch wikipedia entry). – [John](#) Mar 7 '14 at 19:21 ✎
-
- 3 I also saw an explanation that bad news is released in bunches. I have not used GJR-GARCH. I attempted to use multifractal Brownian motion (Mandelbrot) to model asymmetry, but was unable to make it work. – [wcampbell](#) Mar 7 '14 at 21:15
-
- 4 This is at best simplistic. For example, I just took a data set of daily returns on 31 equity indexes. More than half of them have positive skew (using Pearson's skewness) and over 70% are positive on the measure $3 * (\text{mean} - \text{median}) / \text{stdev}$. For commodities you tend to see even more positive skew, as supply and demand shocks can both drive prices up rapidly (e.g. oil, gas and corn in recent years). – [Chris Taylor](#) Mar 8 '14 at 13:01
-

5

Gestational age at delivery (especially for live births) is left skewed. Infants can be born alive very early (although chances of continued survival are small when too early), peak between 36-41 weeks, and drop fast. It is typical for women in the US to be induced if 41/42 weeks, so we don't usually see many deliveries after that point.

answered Jul 30 '18 at 20:20

community wiki
[Sara](#)

4 In fisheries there are often examples of negative skew because of regulatory requirements. For instance the length distribution of fish released in recreational fishery; because there is sometimes a minimum length that a fish must be in order for it to be retained all fish under the limit are discarded. But because people fish where there tends to be legal length fish there tends to be negative skew and mode towards the upper legal limit. The legal length does not represent a hard cut off though. Because of bag limits (or limits on the number of fish that can be brought back to the dock), people will still discard legal size fish when they have caught larger ones.

e.g., Sauls, B. 2012. A Summary of Data on the Size Distribution and Release Condition of Red Snapper Discards from Recreational Fishery Surveys in the Gulf of Mexico. SEDAR31-DW11. SEDAR, North Charleston, SC. 29 pp.

edited Jun 13 '15 at 17:08

community wiki

2 revs

jamesfreinhardt

"Skew towards large sizes" would ordinarily be interpreted as *positive* skew, not "negative." Perhaps you could clarify this answer with an illustration of a typical distribution? The mechanisms you describe--a regulatory upper limit and some tendency to exceed it--could lead either to negative or positive skew, depending on the truncated distribution of the small-size fish (and depending on how the fish are measured: the skewness of their mass distribution would not be the same as the skewness of their length distribution). – whuber ♦ Jun 12 '15 at 16:28

3 Some great suggestions have been made on this thread. On the theme of age-related mortality, machine failure rates are frequently a function of machine age and would fall into this class of distributions. In addition to the financial factors already noted, financial loss functions and distributions typically resemble these shapes, particularly in the case of extreme-valued losses, e.g., as found in BIS III (Bank of International Settlement) estimates of expected shortfall (ES), or in BIS II the value at risk (VAR) as inputs to regulatory requirements for capital reserve allocations.

answered Jun 13 '15 at 17:27

community wiki

Mike Hunter

2 Age of retirement in the U.S. is negatively skewed. The majority of retirees are older with a few retiring relatively young.

edited Mar 1 '18 at 19:46

community wiki

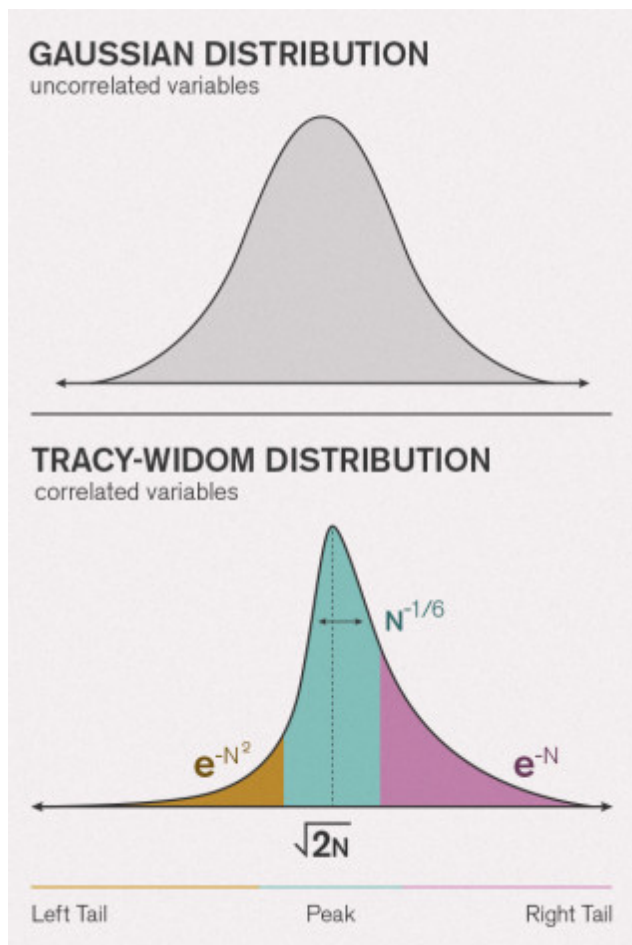
2 revs, 2 users 67%

Ronet Bachman

▲ In random matrix theory, the [Tracy-Widom distribution](#) is right-skewed. This is the distribution of the largest eigenvalue of a random matrix. By symmetry, the smallest eigenvalue has negative Tracy-Widom distribution, and is therefore left-skewed.

2

▼ This is roughly due to the fact that random eigenvalues are akin to charged particles that repel each-other, and hence the largest eigenvalue tends to be pushed away from the rest. Here's an exaggerated picture (taken from [here](#)) :



edited Mar 1 '18 at 21:00

community wiki