TBD

TBD 00 (2025) 1–5

The affine growth of NBA players and other adults

Z.W.T. Mason

Sheffield, UK

Abstract

The heights of NBA players form a skewed distribution rather than follow the normal distribution as the textbooks suggest. There has long been disquiet around whether the normal is the appropriate distribution with some authors using the log-normal instead. For Bernoulli trials offset returns are approximated by the normal distribution, linear returns by the log-normal and it has been recently shown affine returns may be approximated by an appropriately scaled logit-normal distribution. A good fit for the heights of the NBA players is performed by this latter distribution whose parameters are derived using a mixture of Maximum Likelihood Estimation and a grid search. Fitting adult heights of the general population is also undertaken.

1. Introduction

It appears that one of the Keys to remaining a famous Belgian is not to be written out of the Index of history.[1] Adolphe Quetelet saw the normal distribution everywhere in his pursuit of the "average man" and is, perhaps, the reason why the standard textbook example of the normal distribution is adult height. This is not to say that the choice of the normal distribution hasn't been questioned.

Recently Perlman [1] and other students at the University of Dortmund questioned when they were likely to meet someone of negative height. Noting that a probability distribution is the result of some stochastic process, consider the case of a coin tossing game in which even when you lose the toss you still win a set amount only smaller. The result of all these Bernoulli trials is a Binomial distribution over some positive subset of the real line. At scale this can be approximated by a normal distribution whose support is, of course, from minus infinity to plus infinity. i.e. Only in the approximation is there a non-zero probability of a negative result.

As a counter-example to the normal distribution Matthews [2] details John William 'Bud' Rogan (1860s - 1905) of Tennessee who was 2.67 metres tall. Given a mean of 1.70 m and a standard deviation of 0.07 m and assuming the normal distribution this makes him more than 13 standard deviations taller than his fellow American men of the time or one man in 10^{44}. There have been other extremely tall men since so this is not that unusual. A contemporary of Mr Rogan was Charles Sherwood Stratton (1838 – 1883) who, at 0.64 m tall, was 15 standard deviations shorter than the mean. Again, given that the continuous distribution may be an approximation, the question of whether outliers, no matter how frequent they are, are significant has to be questioned.

Alternatives to the normal distribution have been tried, in particular the log-normal distribution. For example, Yuan [3] fitted the heights of Glaswegian schoolboys with some success apart from a few outliers. Perhaps it is worthwhile to look at whether some other function of the random variable is normally distributed. By way of example, if there were some upper scale applicable to humans then the logit-normal distribution may be appropriate. Matching the first

[1] Ancel Keys renamed the Quetelet Index to the Body Mass Index.

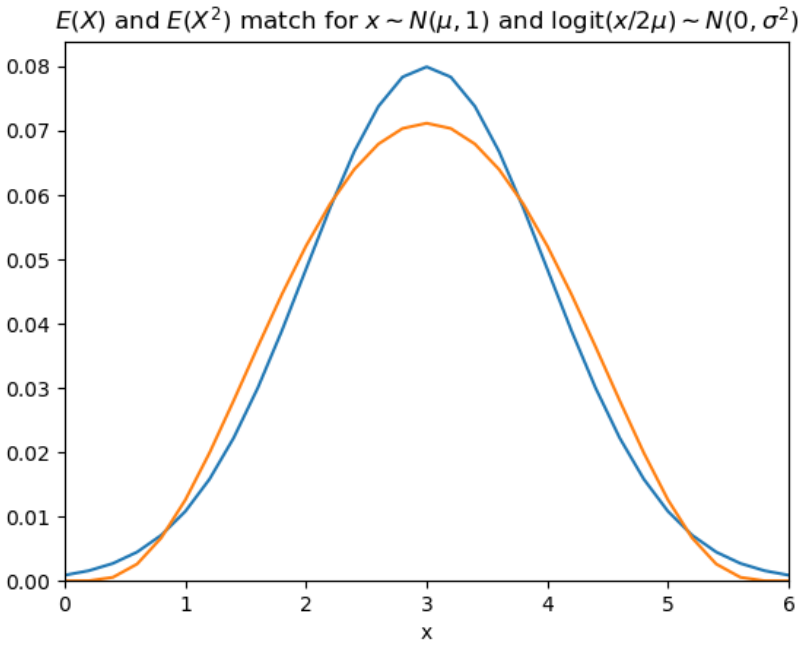

Figure 1. Comparison of normal and logit-normal

and second moments between normal and symmetric logit-normal distributions produces very similar curves as can be seen in figure 1. The logit-normal distribution is much overlooked but important, for example it moderates the growth of a well-mixed epidemic from exponential to logistic as shown by the author [4].

The author [5] also showed, for stock prices, that affine returns on coin tosses can be approximated by a stretched-out logit-normal distribution for the case when the growth converges to a finite support. The same shapes were also evident in the case where the support grows exponentially with t when $\beta > 1$ in equation 1.

$$\left(0, \delta \frac{1-\beta^t}{1-\beta}\right) \tag{1}$$

The process itself being described by the random variable S_t growing from $S_0 = 0$ with probability p

$$S_{t+1} = \beta S_t + \delta,\ 0 < \beta \neq 1,\ \delta > 0 \tag{2}$$

and shrinking with probability $q = 1 - p$

$$S_{t+1} = \alpha S_t,\ 0 < \alpha < 1 \tag{3}$$

Should this model be appropriate then it can be easily explained to children that every day someone tosses a coin for them. Heads they grow, tails they shrink, and that, if they have a run of bad luck, then they will shrink away to nothing. Whether the model is appropriate though hinges on whether it fits the data.

2. Parameter estimation

The Maximum Likelihood Estimators for the log-normal and logit-normal distributions are the same as for the normal distribution fitted to the values transformed by the appropriate function, so the fitting is done in a highly distorted space. The probability density function for the scaled logit-normal is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \frac{L}{x(L-x)} \exp\left(-\frac{(\log(x) - \log(L-x) - \mu)^2}{2\sigma^2}\right) \tag{4}$$

Taking logarithms

$$-\log(\sigma) - \frac{1}{2}\log(2\pi) + \log(L) - \log(x) - \log(L-x) - \frac{(\log(x) - \log(L-x) - \mu)^2}{2\sigma^2} \tag{5}$$

Differentiating with respect to μ, σ, L results in, respectively

$$\frac{\log(x) - \log(L-x) - \mu}{\sigma^2} \tag{6}$$

$$-\frac{1}{\sigma} + \frac{(\log(x) - \log(L-x) - \mu)^2}{\sigma^3} \tag{7}$$

$$\frac{1}{L} - \frac{1}{L-x} + \frac{\log(x) - \log(L-x) - \mu}{\sigma^2} \frac{1}{L-x} \tag{8}$$

Resulting in the usual, if L were equal to one, estimates of

$$\mu\Sigma_i 1 = \Sigma_i (\log(x_i) - \log(L - x_i)) \tag{9}$$

$$\sigma^2\Sigma_i 1 = \Sigma_i (\log(x_i) - \log(L - x_i) - \mu)^2 \tag{10}$$

The final derivative was not summed over the data and used to estimate L but was used to estimate the error. The mean square error between the formula and the histogram was also calculated. Finally, the result was checked to see whether it could be improved upon by searching through a grid of parameters.

3. Data analysis

3.1. Data for NBA players

Biometric, biographic and basic box score stats from 1996 to 2022 season for NBA players was downloaded from Kaggle [6]. Maximum Likelihood Estimation was performed with L from 232 to 1000 cm with the mean square error increasing at each step. Subsequently a matrix of tests was performed with L from 231.2 to 1000 cm, μ from -3 to 3 and σ^2 from 0.001 to 0.5. A minimum was observed for $L = 231.344\,\text{cm}$, $\mu = 1.9$, $\sigma^2 = 0.14072$ which is shown in figure 2. The heights are not normally distributed and the fit seems reasonable.

3.2. Data from the CDC NHANES

2007-2008 questionaire data was downloaded from the Continuous National Health and Nutritional Examination Survey conducted for the US CDC [7]. The heights were then split into separate sets for men and women.

For men the Maximum Likelihood Estimation was performed with L from 82 to 180 inches with the mean square error decreasing to a minimum at $L = 113\,\text{in}$, $\mu = 0.459$, $\sigma^2 = 0.015$. Subsequently a matrix of tests was performed with a minimum at $L = 188.2\,\text{in}$, $\mu = -0.54$, $\sigma^2 = 0.006$ which is shown in figure 3. The heights do not seem to be skewed and the fit seems reasonable.

For women the Maximum Likelihood Estimation was performed with L from 80 to 1000 inches with the mean square error decreasing monotonically. The fit for women of $L = 233\,\text{in}$, $\mu = -0.98$, $\sigma^2 = 0.004$ was indistinguishable from the fit at $L = 1000\,\text{in}$, $\mu = -2.69$, $\sigma^2 = 0.0024$. Subsequently a matrix of tests was performed with a minimum at $L = 149.6\,\text{in}$, $\mu = -0.3$, $\sigma^2 = 0.006$ which is shown in figure 4. The heights are not normally distributed and the fit seems reasonable.

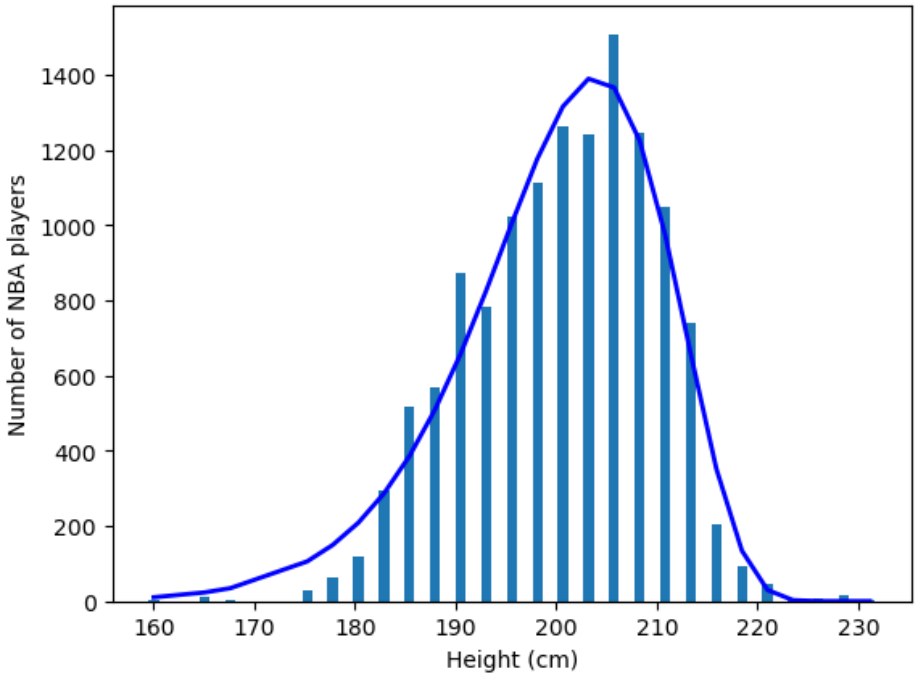

Figure 2. Fit of NBA player heights

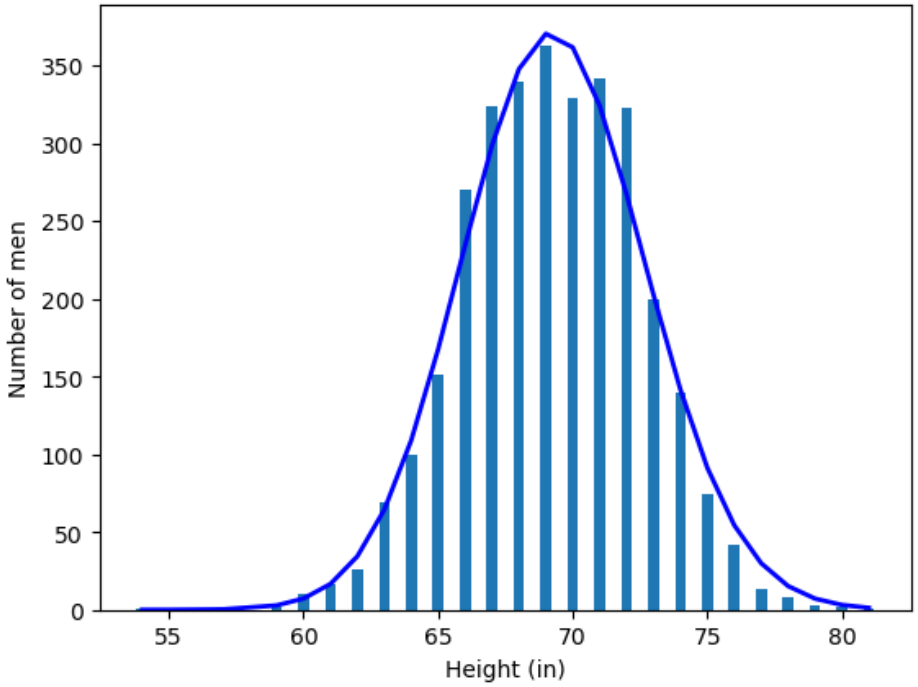

Figure 3. Fit of American male heights

4. Discussion

The heights of NBA players were seen to be especially well fit by the logit-normal distribution due to the degree of skewness. When it came to the general population the data seemed to be better matched to the normal distribu-

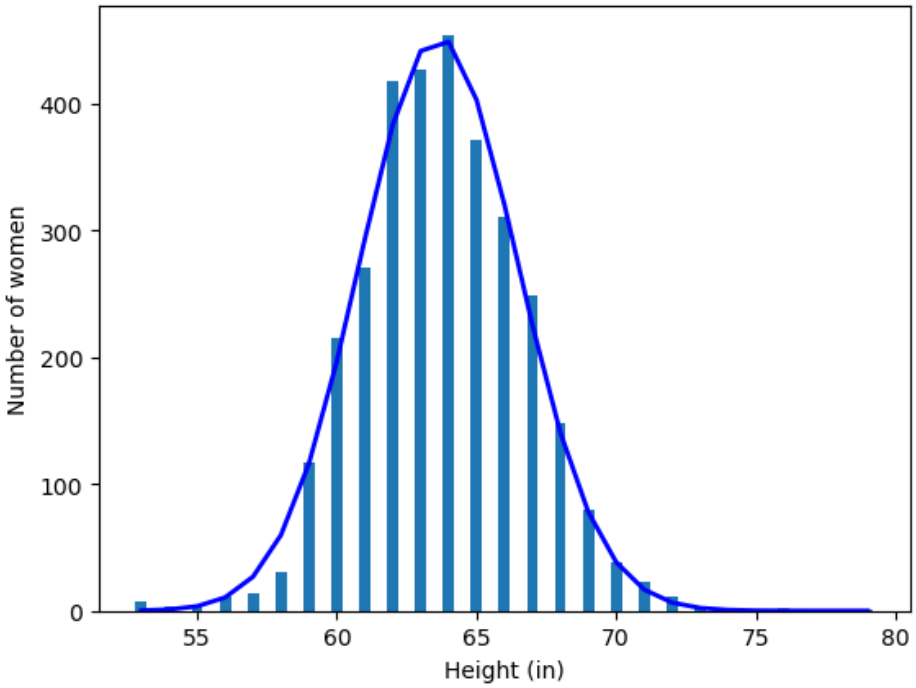

Figure 4. Fit of American female heights

tion, however the logit-normal distribution still made a good fit. This was probably because its parameters can be manipulated to nearly match the normal distribution away from its tails.

That the Maximum Likelihood Estimation wasn't sufficient to determine the parameters may be due to the support being semi-infinite. i.e. The length scale tends to infinity. A rough grid search, although ugly and a misuse of cheap computing power, seemed to be sufficient.

The length scale for modern American men at 188 inches was seen to be comfortably larger than the height of Mr Rogan at 105 inches, however the corresponding scale for NBA players was only 91 inches. In general, this length scale seems to be rather arbitrary.

References

[1] P. Perlman, When will we see people of negative height?, Significance 10 (1) (2013) 46–48. arXiv:https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1740-9713.2013.00642.x, doi:https://doi.org/10.1111/j.1740-9713.2013.00642.x. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1740-9713.2013.00642.x

[2] R. Matthews, Chancing It: The Laws of Chance and How They Can Work for You, Profile Books, London, 2016.

[3] P. T. Yuan, On the logarithmic frequency distribution and the semi-logarithmic correlation surface, The Annals of Mathematical Statistics 4 (1) (1933) 30–74. doi:10.1214/aoms/1177732821.

[4] Z. Mason, Naturally extending the standard sir model to stochastic growth, TBDdoi:10.31219/osf.io/y6ckv.

[5] Z. Mason, Affine returns on bernoulli trials in finance, TBD.

[6] J. Cirtautas, Nba players, https://www.kaggle.com/datasets/justinas/nba-players-data (2023).

[7] 2007-2008 questionnaire data - continuous nhanes, https://wwwn.cdc.gov/nchs/nhanes/search/datapage.aspx (2008).