Central limit theorem and large deviation theory

Pick a probability distribution for a random variable x that takes a value from a discrete set. Define

$$z^{(N)} = \sum_{n=1}^{N} x_n,$$

where $\{x_n\}$ are independent draws from the distribution. What is the probability distribution of $z^{(N)}$?

I recommend investigating this question on a computer. Here's example code in `octave`.

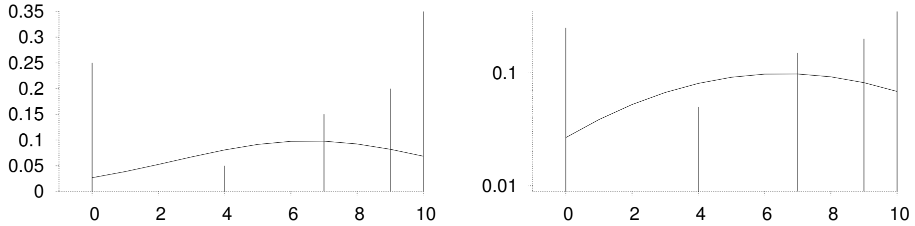
```
p = [ 0.25, 0,  0,  0,  0.05,  0,  0,  0.15,  0,  0.2,  0.35 ];
y = p ;
gplot y' w impulse  1;
y = conv( y , p ) ; # Find the probability of x1+x2 by convolution
gplot y' w impulse  1;
y = conv( y , p ) ; # Find the probability of x1+x2+x3
for N=3:200
  gplot y' w impulse  1;
  y = conv( y , p ) ; # Find the probability of x1+x2 by convolution
endfor
```

You'll find that once you add about $N = 20$ of the random variables, no matter what their original distribution `p`, the probability distribution looks amazingly close to a Gaussian.

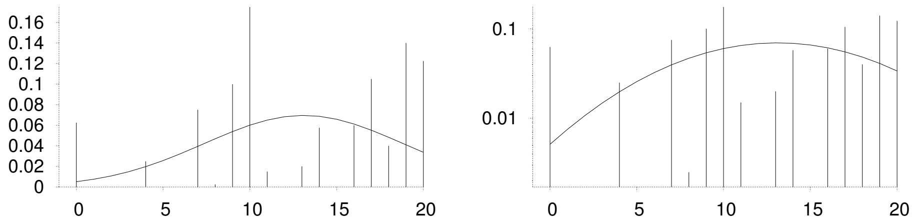
The Gaussian becomes narrower, relative to the range of possible values of $z^{(N)}$, as N increases. The probabilities of extreme values of $z^{(N)}$ are not well approximated by a Gaussian; but since the Gaussian itself is getting narrower as N increases, the quality of the Gaussian approximation gets better in the sense that it is a good approximation out to an increasingly large number of standard deviations.

In my figures the vertical lines show the true probability distribution, and the curve shows a Gaussian approximation.

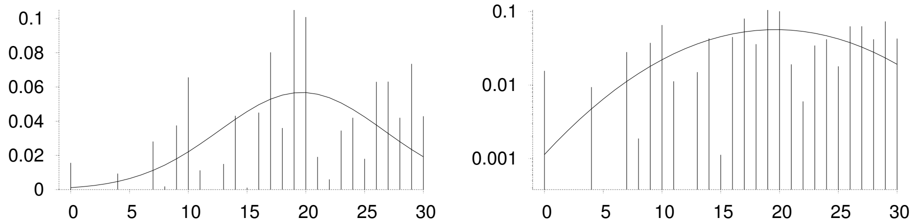
1


0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0 2 4 6 8 10
0.1
0.01
0 2 4 6 8 10

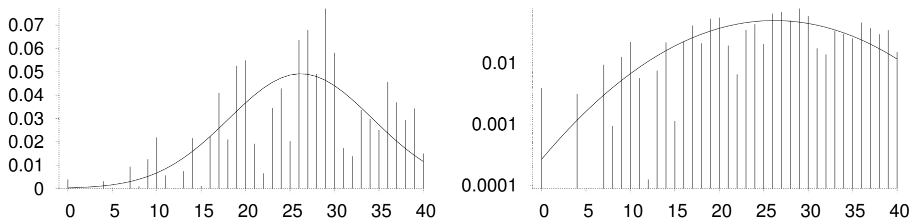
2


0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
0 5 10 15 20
0.1
0.01
0 5 10 15 20

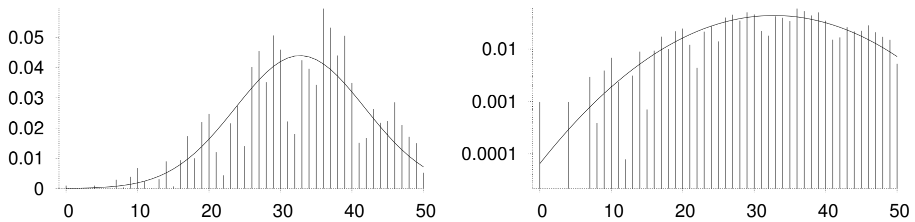
3


0.1
0.08
0.06
0.04
0.02
0
0 5 10 15 20 25 30
0.1
0.01
0.001
0 5 10 15 20 25 30

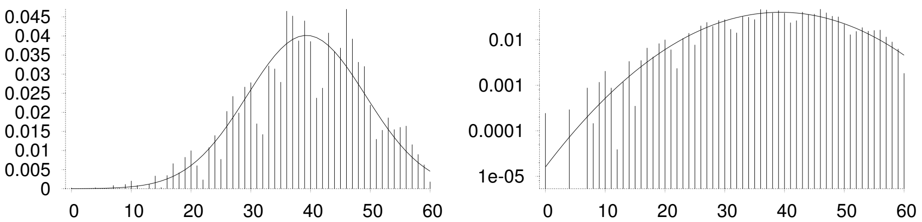
4


0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
0 5 10 15 20 25 30 35 40
0.01
0.001
0.0001
0 5 10 15 20 25 30 35 40

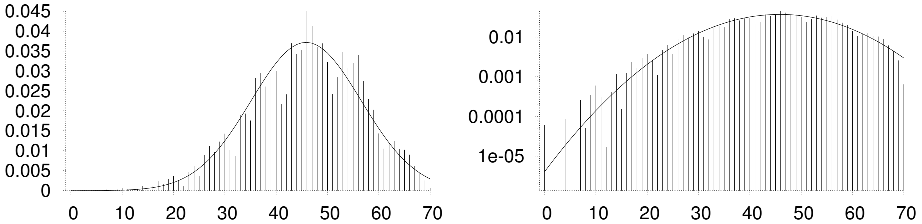
5


0.05
0.04
0.03
0.02
0.01
0
0 10 20 30 40 50
0.01
0.001
0.0001
0 10 20 30 40 50

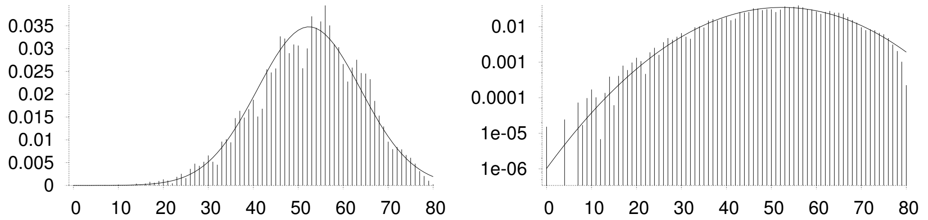
6


0.045
0.04
0.035
0.03
0.025
0.02
0.015
0.01
0.005
0
0 10 20 30 40 50 60
0.01
0.001
0.0001
1e-05
0 10 20 30 40 50 60

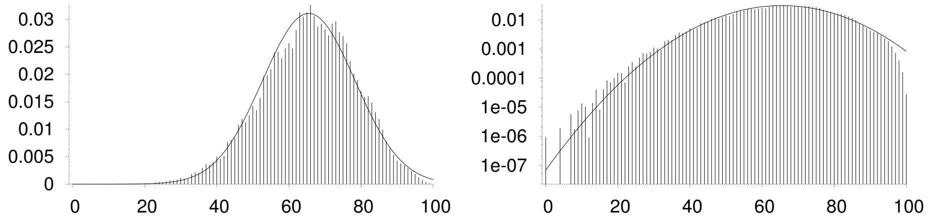
7


0.045
0.04
0.035
0.03
0.025
0.02
0.015
0.01
0.005
0
0 10 20 30 40 50 60 70
0.01
0.001
0.0001
1e-05
0 10 20 30 40 50 60 70

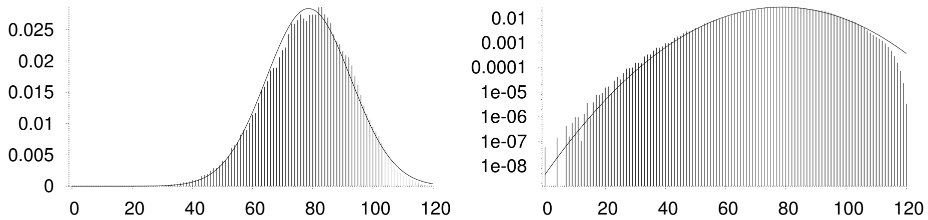
8



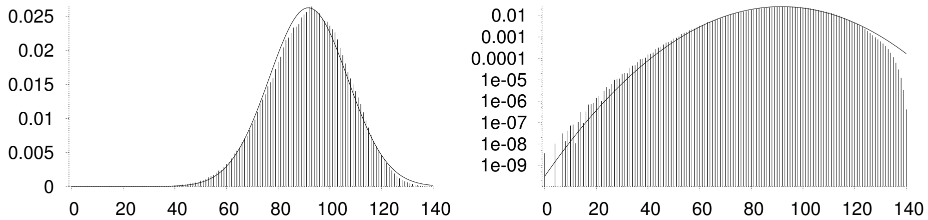
10



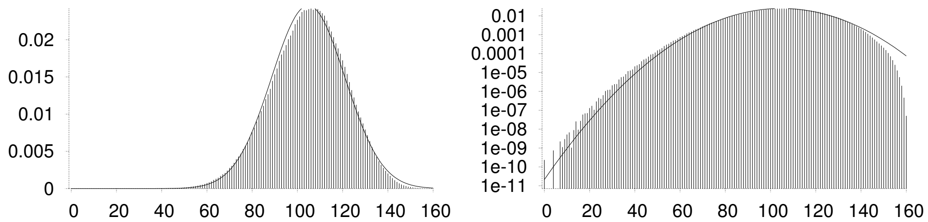
12



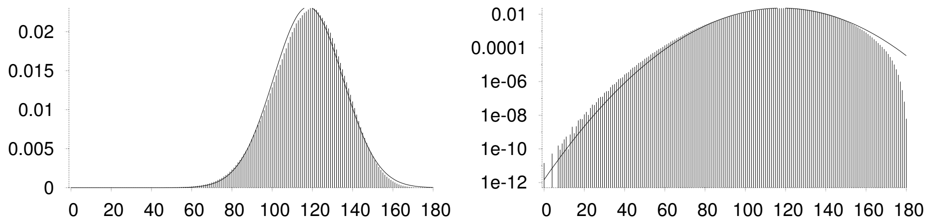
14



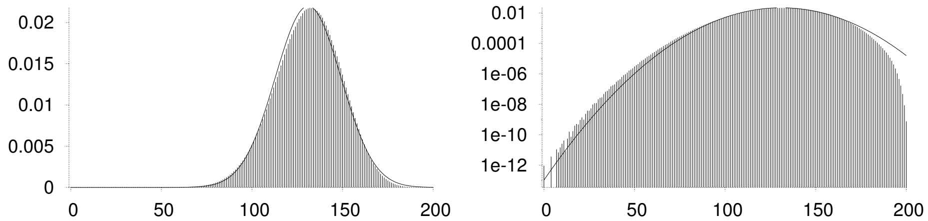
16



18

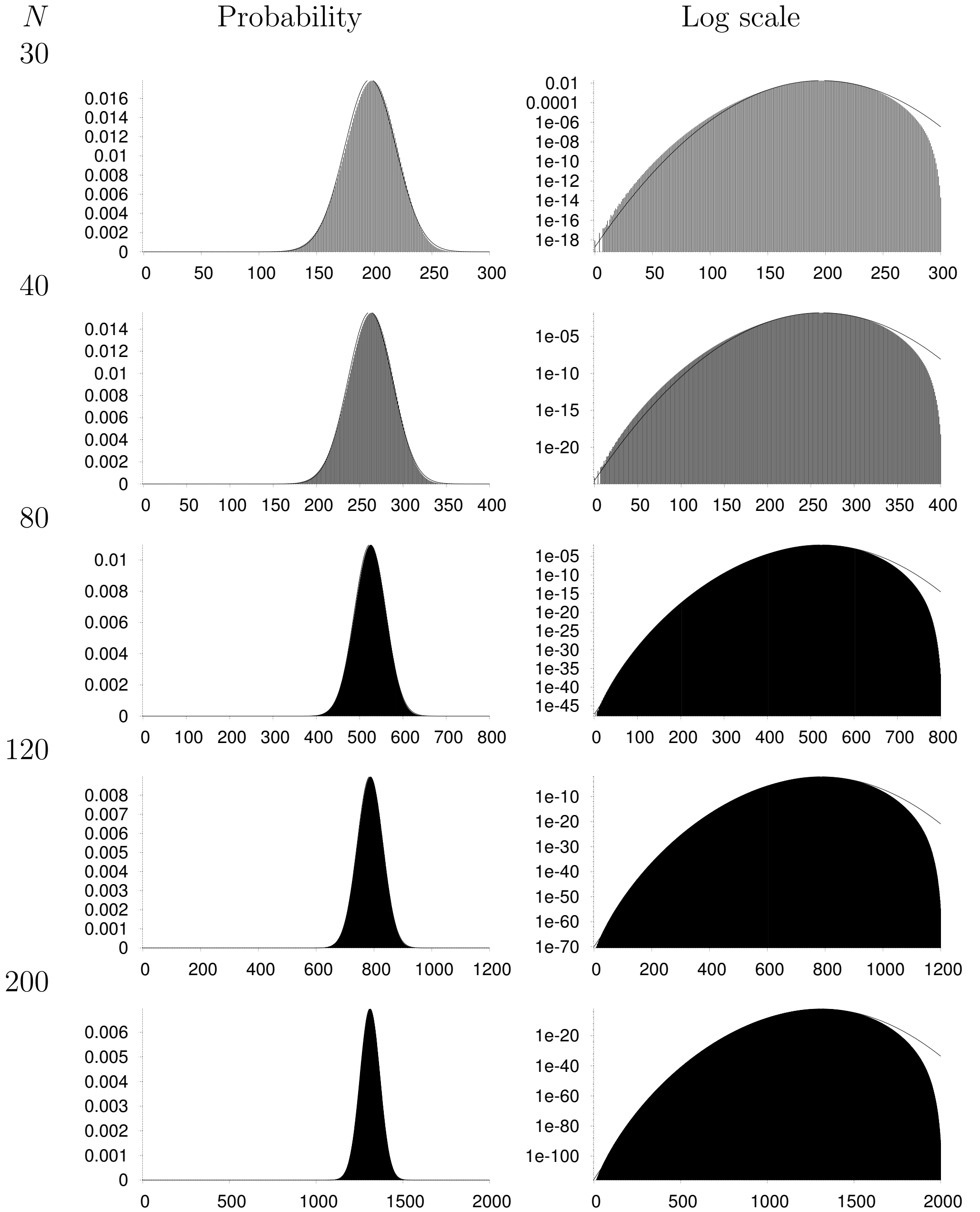


20



Large deviations

If we plot the probability distribution of $z^{(N)}$ on a logarithmic scale and look at the probability distribution for extreme values of $z^{(N)}$, we discover a striking result: if we set the horizontal range to the full range $[z_{\min}, z_{\max}]$, and the vertical scale to the full range of probabilities, then *the whole picture's appearance becomes independent of* N.



The strange lop-sided log-scale envelope is determined by the original probability distribution. The envelope tends to a concave function known as the *large deviation rate function* of the random variable.

Large deviation theory is important in several fields.

- Actuaries use large deviation rate functions to calculate safe premiums. What is the probability that the sum of all the insurance claims in one year will exceed the insurance company's revenue?
- Communications engineers use large deviation rate-functions to calculate safe buffer sizes for network traffic.

There is a strong connection between large deviation theory and the entropy function of statistical physics.

One of the research groups working on large deviation theory is at DIAS in Ireland: `http://www.stp.dias.ie/APG/poster.html`.

Notes

An `octave` program that displays graphs like the ones in this paper is `convolve.m` available from `http://www.aims.ac.za/~mackay/octave/centrallimit/`. The `octave` program that made the figures for this paper is `convolveps.m` available from the same place.