Extreme Value Distributions

Consider an experiment where you have sampled n items but you are not interested in the mean or variance of these n items. Rather you want to know only the largest of the n values.

Equivalently consider the question of the largest flood in a century. This samples 100 yearly flood values but is concerned with only the largest.

Extreme Value Distributions

More formally ...

Assume n values are sampled from a population; x_i where $i = 1, \ldots, n$. Order these values such that

$$x_1 > x_2 > \cdots > x_n$$

What is the distribution of x_1?

Extreme Value Distributions

More formally ...

Assume n values are sampled from a population; x_i where $i = 1, \ldots, n$. Order these values such that

$$x_1 > x_2 > \cdots > x_n$$

What is the distribution of x_1?

Doesn't this depend on the distribution of x?

Extreme Value Distributions

More formally ...

Assume n values are sampled from a population; x_i where $i = 1, \ldots, n$. Order these values such that

$$x_1 > x_2 > \cdots > x_n$$

What is the distribution of x_1?

Doesn't this depend on the distribution of x? No!

Extreme Value Distributions

Given **any*** distribution, $F(x)$ for the x_i values, the distribution of x_1 will be

$$G(x) = e^{-e^{-nF'(x_{on})(x-x_{on})}}$$

(where x_{on} is the value that has expectation of being exceeded once).

This is known as a type I or Gumbel extreme value distribution (for unbounded random variables).

* with some qualifications.

Extreme Value Distributions

Basically any time you see a double exponential it is likely to be a distribution for an extreme value.

These distributions are commonly used in reliability engineering, in weather forecasting, in survival analysis and so on.

Extreme Value Distributions - Examples

The parameters μ and $\beta > 0$ measure 'location' and 'scale' respectively.

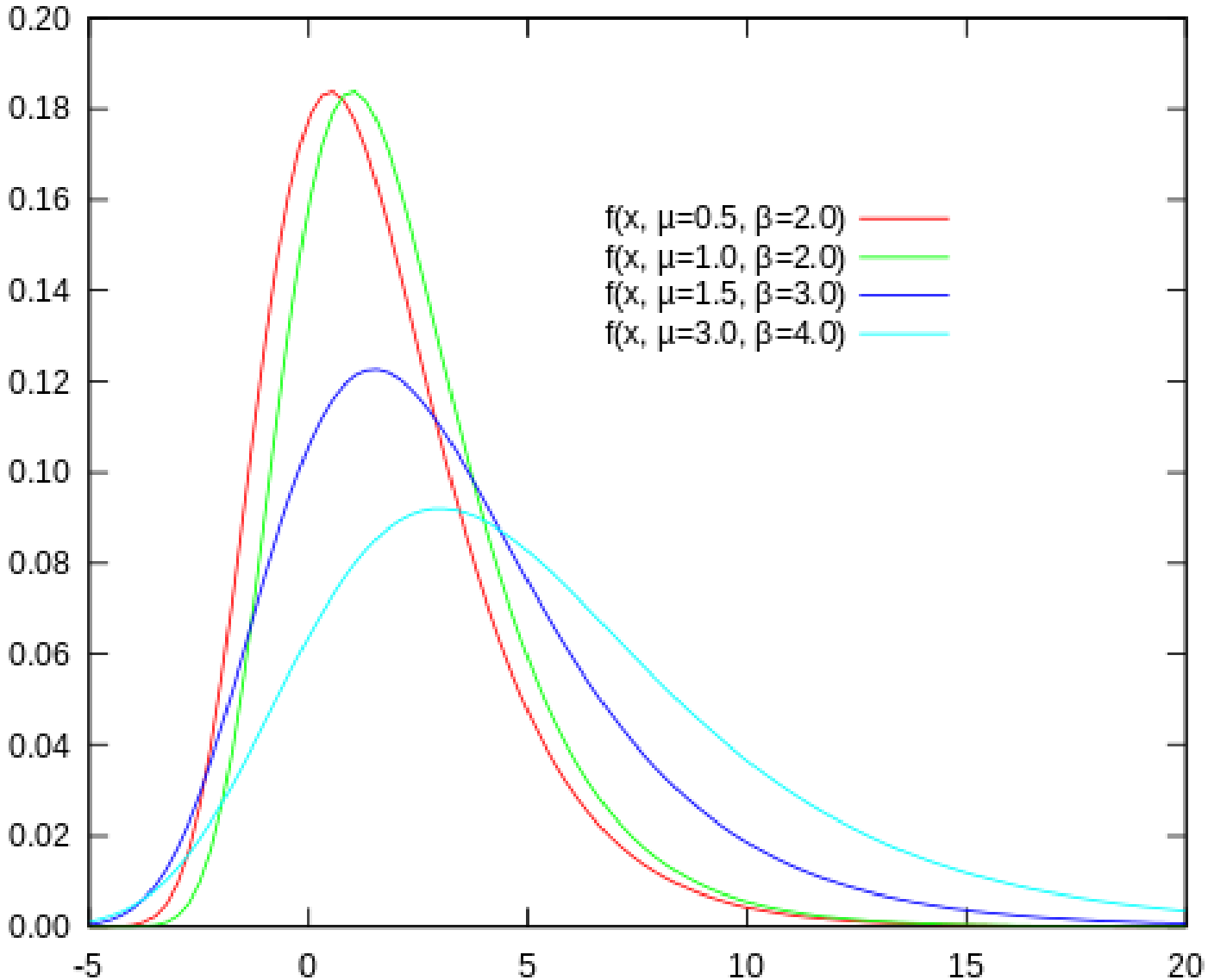
The general probability density function (PDF) is,

$$\frac{1}{\beta} e^{-(z+e^{-z})} \text{ where } z = \frac{x-\mu}{\beta}$$

The standard Gumbel is with $\mu = 0$ and $\beta = 1$, so the PDF is

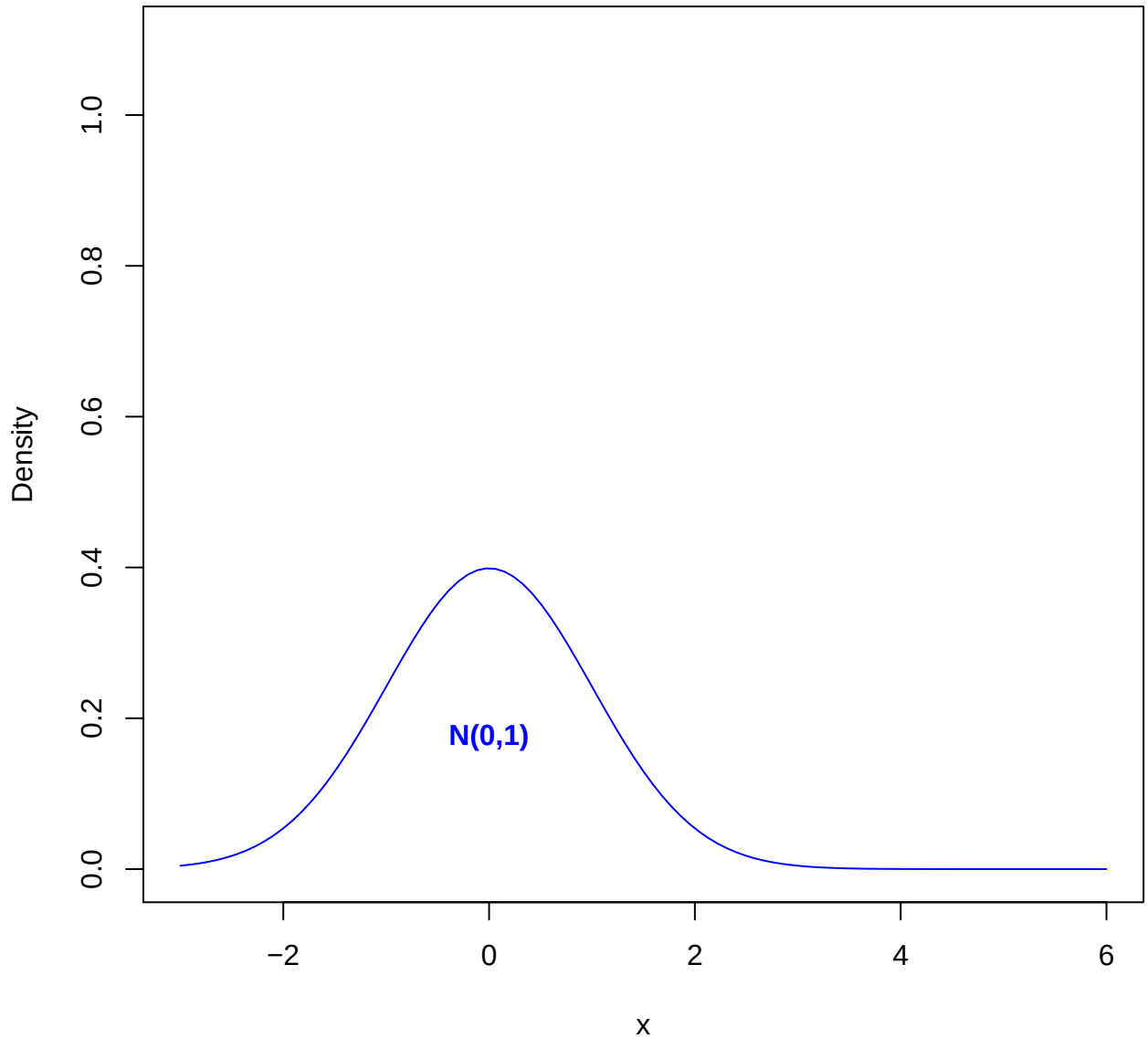
$$e^{-(x+e^{-x})}$$

Extreme Value Distributions - Examples

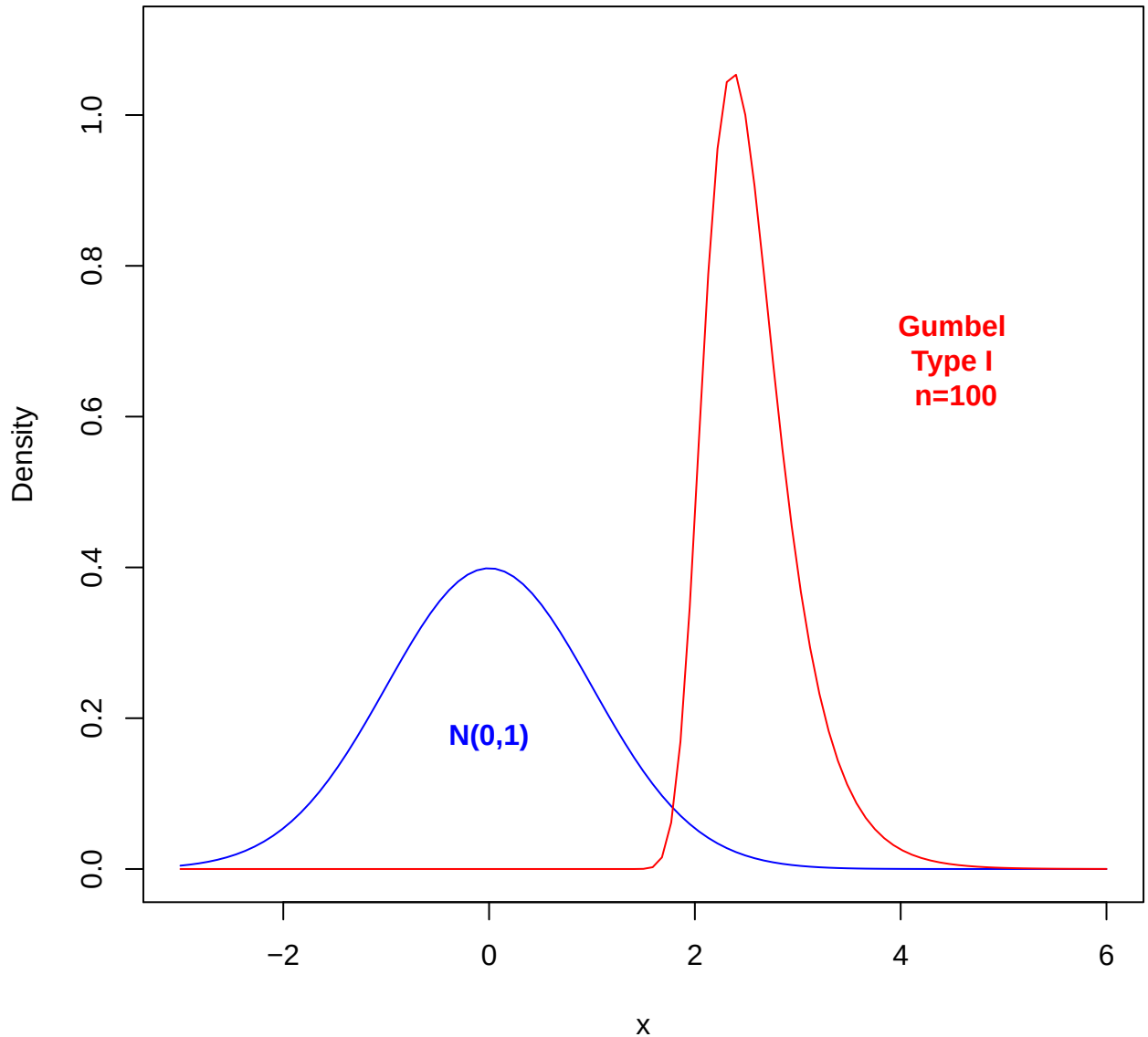


Wikipedia: http://en.wikipedia.org/wiki/Gumbel_distribution

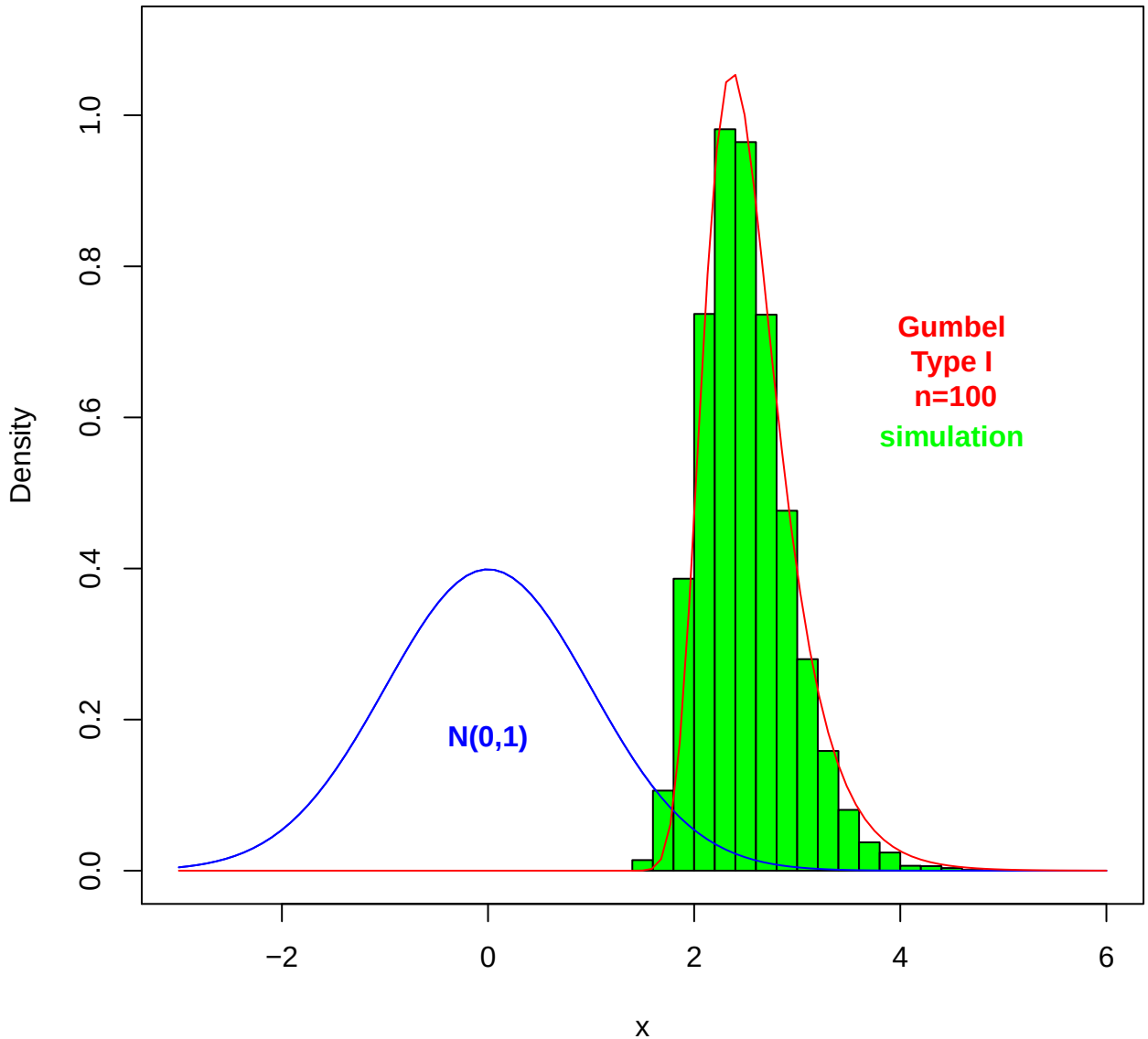
Extreme Value Distributions - Examples



Extreme Value Distributions - Examples



Extreme Value Distributions - Examples



For ungapped BLAST like searches

The probability of a database match score larger than x arising by chance alone is therefore

$$P(s \geq x) = 1 - e^{-e^{-\lambda(x-u)}}$$

where for ungapped alignments

$$u = \frac{\ln(Kmn)}{\lambda}$$

and m, n are the lengths of the query and library sequence and K and λ are constants that depend on the substitution scores and the sequence compositions.

For ungapped BLAST like searches

This formula can be rewritten as:

$$1 - e^{-Kmn(e^{-\lambda x})}$$

which shows that the probability of observing larger scores for unrelated library sequences increases logarithmically with the length of the library sequence.

(Pearson - FASTA documentation).

For ungapped BLAST like searches

The main point of all of this is to arrive at an expression for the expected number of HSPs having a score equal to or larger than S.

It is

$$E = Kmne^{-\lambda S}$$

Here E is called an expect value. Again, it gives an estimate of the **number** of matches that would have a score of S (or better) by chance alone.

For ungapped BLAST like searches

Note that E is **NOT** a probability. It is however related to the probability of getting a score as good as S or better (at least in theory).

Specifically,

$$P = 1 - e^{-E}$$

For ungapped BLAST like searches

To make the confusion even worse . . . when E is small it is close to P.

E	P
10	0.99995
5	0.99326
2	0.86466
1	0.63212
0.1	0.09516
0.05	0.04877
0.001	0.00099
0.0001	0.00010

Never-the-less, beware that E is **not** a probability.

For BLAST like searches

Because not all of the assumptions can be met for this theory to hold accurately, the E values are often adjusted to be more conservative before declaring them to be significant.

Some suggest, dividing the significance level by the number of database comparisons (e.g. 10^6). Others suggest more ad-hoc methods of considering an E value significant if it is less than 10^{-5} or 10^{-6}.

There is no strict consensus.