A reference for probability lingo in the Track and Field Case Study

## Random sampling in the Track and Field case study

**What does it mean that you are drawing or sampling from a distribution?**

Imagine that you had a bunch of runners. We stacked them up in separate bins of some measurement of performance level. Now if we go back and choose a random square from that distribution, we read the score, record it as another entry in our vector, and put it back. If you do this 10 times, you get a vector of length 10. If you do it 100 times, you get a vector of length 100, etc. The computer simulates this process. If you use rnorm(10) it will do it 10 times, pulling from a “**normal distribution**” bell-shaped curve with mean 0 and standard deviation 1 (means that about 2/3rd of your numbers will be between -1 and 1, with most near 0).

**How does the assumption about distribution type affect the maximum of a sample?**

Imagine you had the kind of population that was mostly between 0 and 1, and then only a couple at a 2 or a 3. You'd mostly be drawing 0 and 1s and rarely hit a 3 or greater. Contrast that with a distribution that has all values equally likely between 0 and 4. Getting a 3 or more is no longer a rare event. So the underlying distribution you assume, matters.

**How does sample size affect the maximum of that sample?**

In the case of a uniform distribution, where all values between 0 and 4 are equally likely, there might be little effect. The maximum of each of the samples is probably similar no matter how many are drawn. However, in the case of the distributions considered in the project, getting a 3 or larger was a rare event. So if you have a small sample size, you are unlikely to experience that event - the maximum is likely to be only very high when your sample size is very large.

## Continuous Distributions Quick Reference Guide

This is a [full reference list of probability distributions](https://en.wikipedia.org/wiki/List_of_probability_distributions). Below is a list of continuous distributions used in this case study and why we might consider using each:

**Uniform distribution**: [Wikipedia reference](https://en.wikipedia.org/wiki/Uniform_distribution_%28continuous%29)

You would use the uniform distribution if you assumed that all outcomes (within some constraint of a<X< b) are equally likely.

In R, to draw random samples from a uniform continuous distribution, use runif().



**Normal distribution**: [Wikipedia reference](https://en.wikipedia.org/wiki/Normal_distribution)

This is the classic “bell-shaped” distribution. You would use the normal distribution if you believe that the mean, 𝜇, = the median = the mode.

In R, to draw random samples from a continuous distribution, use rnorm().

**Folded normal**: [Wikipedia reference](https://en.wikipedia.org/wiki/Folded_normal_distribution) 

This distribution assumes you only consider the magnitude of a number normally distributed, and not the sign. Mathematically, this is equivalent to taking the absolute value of the normal distribution.

In R, to draw random samples from a continuous distribution, use abs(rnorm()).

**Exponential distribution**: [Wikipedia reference](https://en.wikipedia.org/wiki/Exponential_distribution)

An exponential distribution typically emerges when looking at the waiting time for an event to occur.

In R, to draw random samples from a continuous distribution, use rexp().



**Log-normal distribution:** [Wikipedia reference](https://en.wikipedia.org/wiki/Log-normal_distribution)

This may be used for a variable that is described as the product of many variables or for variables that are measured on a log scale (such as pH or seismic activity).

In R, to draw random samples from a continuous distribution, use rlnorm().