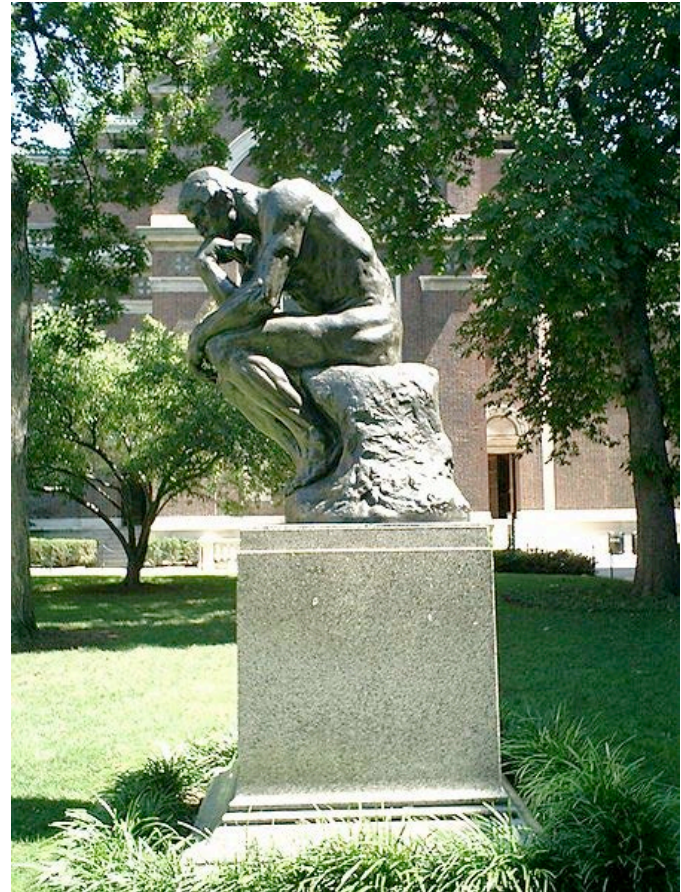


Sampling and Standard Error

What Can We Conclude from 1 Sample?

- More than you might think
- Thanks to the Central Limit Theorem



Recall Central Limit Theorem

- Given a sufficiently large sample:
 - 1) The means of the samples in a set of samples (the sample means) will be approximately normally distributed,
 - 2) This normal distribution will have a mean close to the mean the population, and
 - 3) The variance of the sample means will be close to the variance of the population divided by the sample size.
- Time to use the 3rd feature
- Compute *standard error of the mean* (SEM or SE)

Standard Error of the Mean

$$SE = \frac{\sigma}{\sqrt{n}}$$

- Does it work?

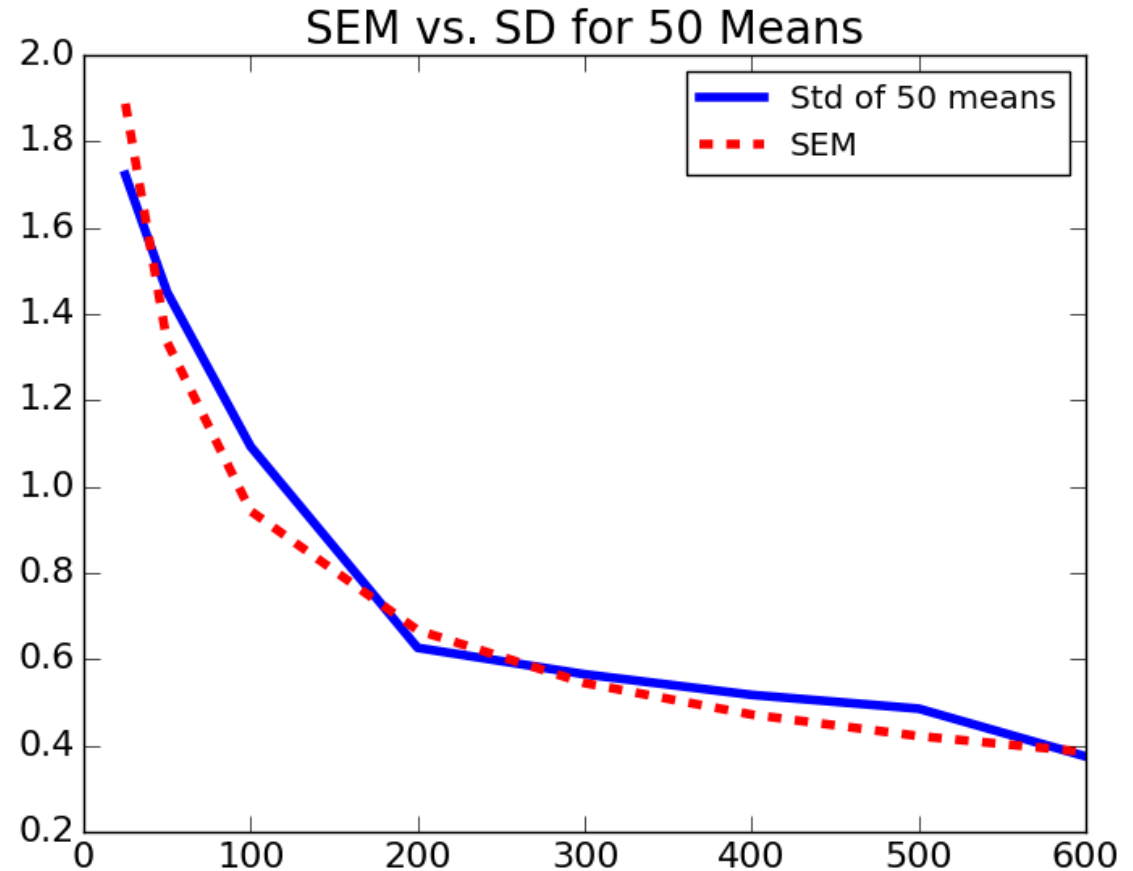
Testing the SEM

```
sampleSizes = (25, 50, 100, 200, 300, 400, 500, 600)
numTrials = 50
population = getHighs()
popSD = numpy.std(population)
sems = []
sampleSDs = []
for size in sampleSizes:
    sems.append(sem(popSD, size))
    means = []
    for t in range(numTrials):
        sample = random.sample(population, size)
        means.append(sum(sample)/len(sample))
    sampleSDs.append(numpy.std(means))
pylab.plot(sampleSizes, sampleSDs,
            label = 'Std of 50 means')
pylab.plot(sampleSizes, sems, 'r--', label = 'SEM')
pylab.title('SEM vs. SD for 50 Means')
pylab.legend()
```

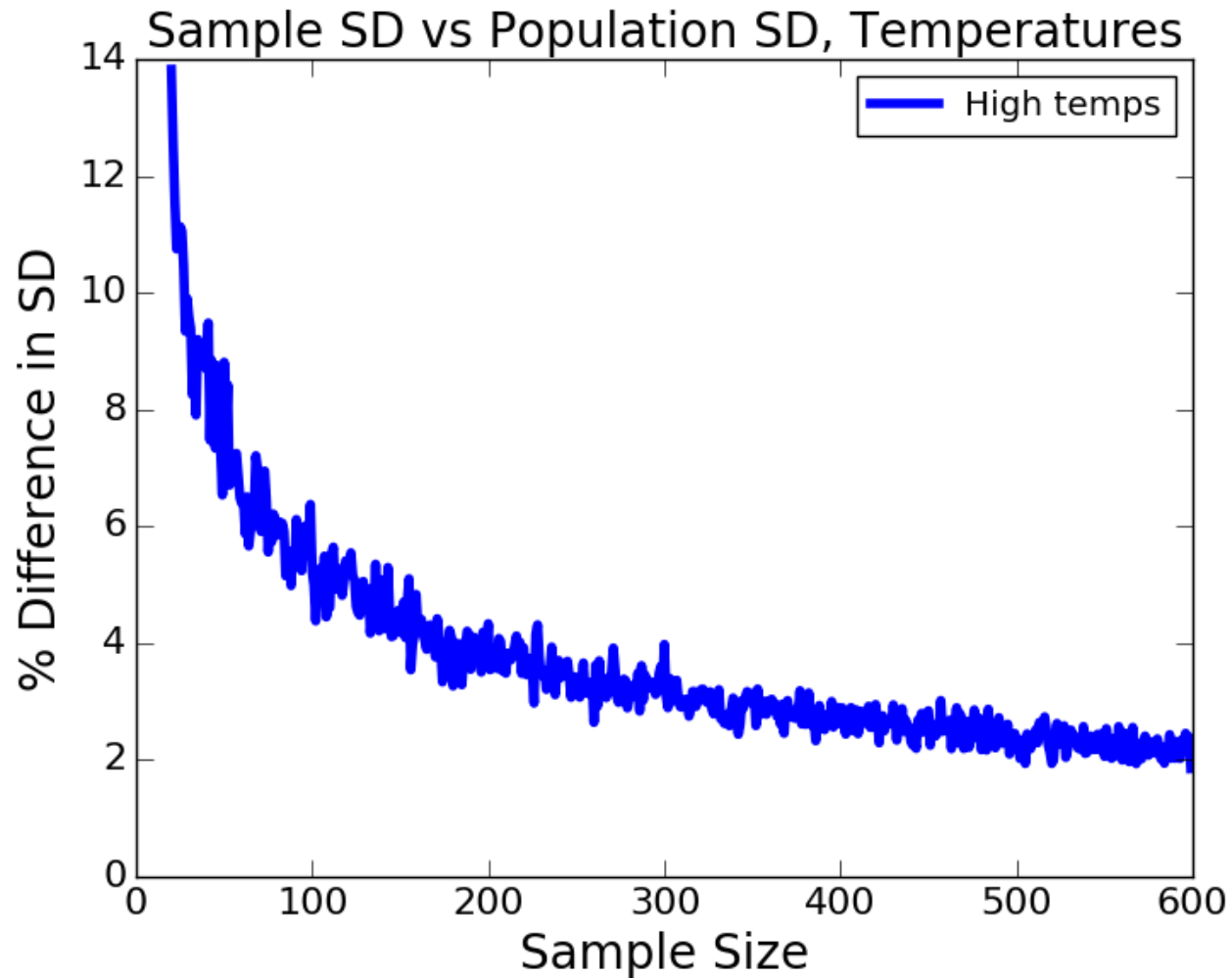
Standard Error of the Mean

$$SE = \frac{\sigma}{\sqrt{n}}$$

But, we don't
know standard
deviation of
population

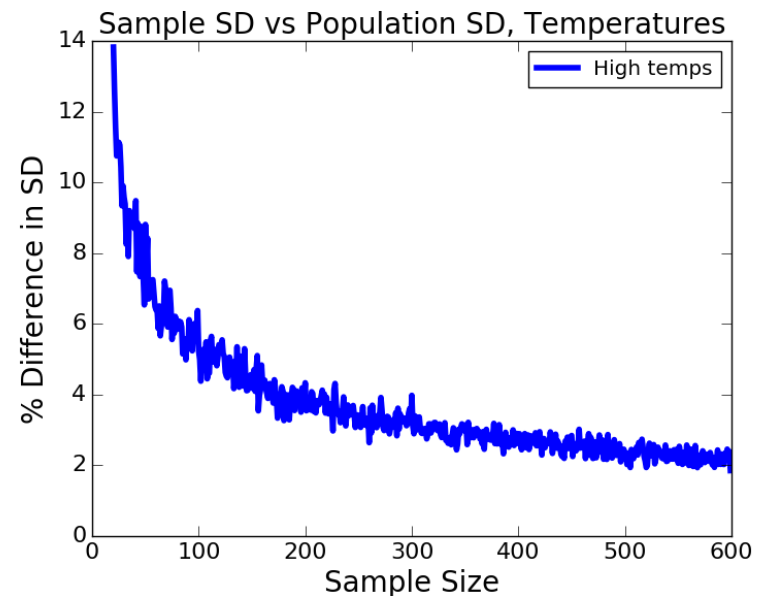


Sample SD vs. Population SD



The Point

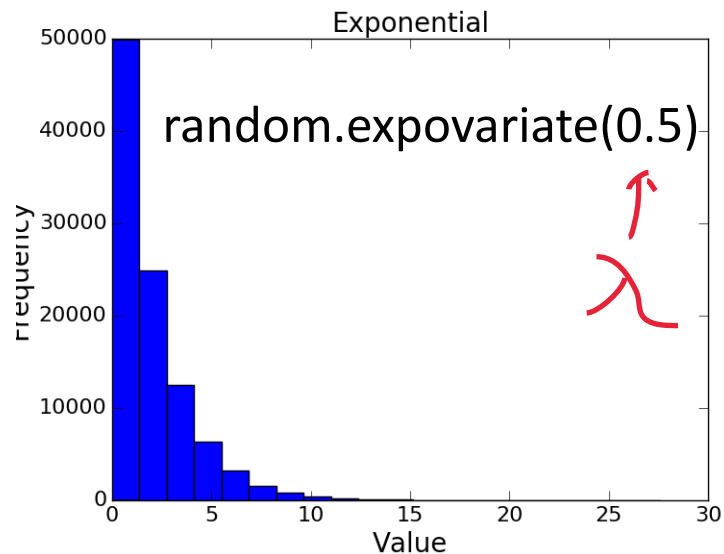
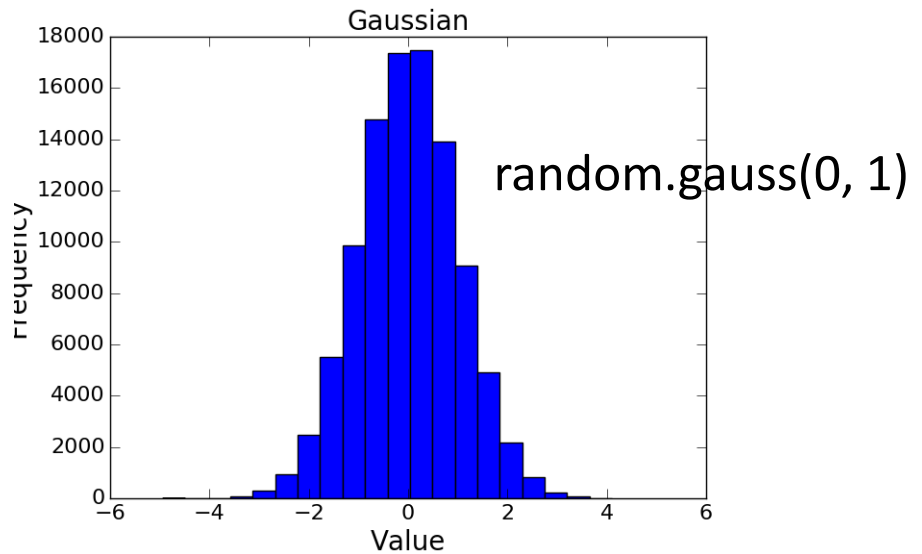
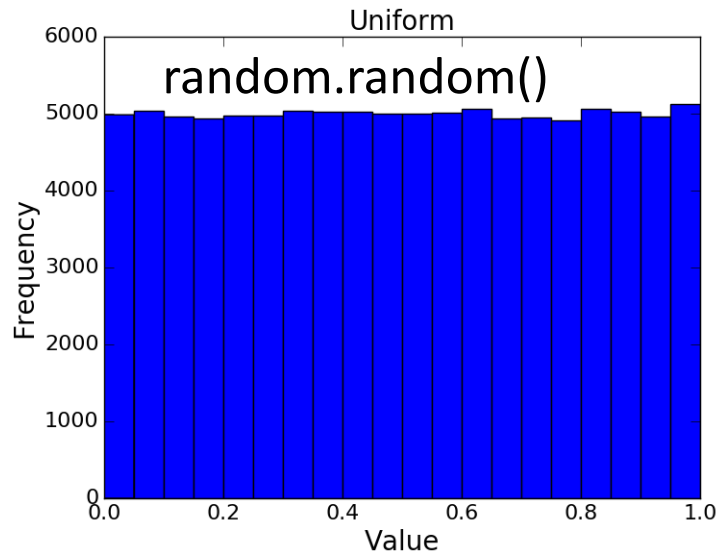
- Once sample reaches a reasonable size, sample standard deviation is a pretty good approximation to population standard deviation
- True only for this example?
 - Distribution of population?
 - Size of population?



Looking at Distributions

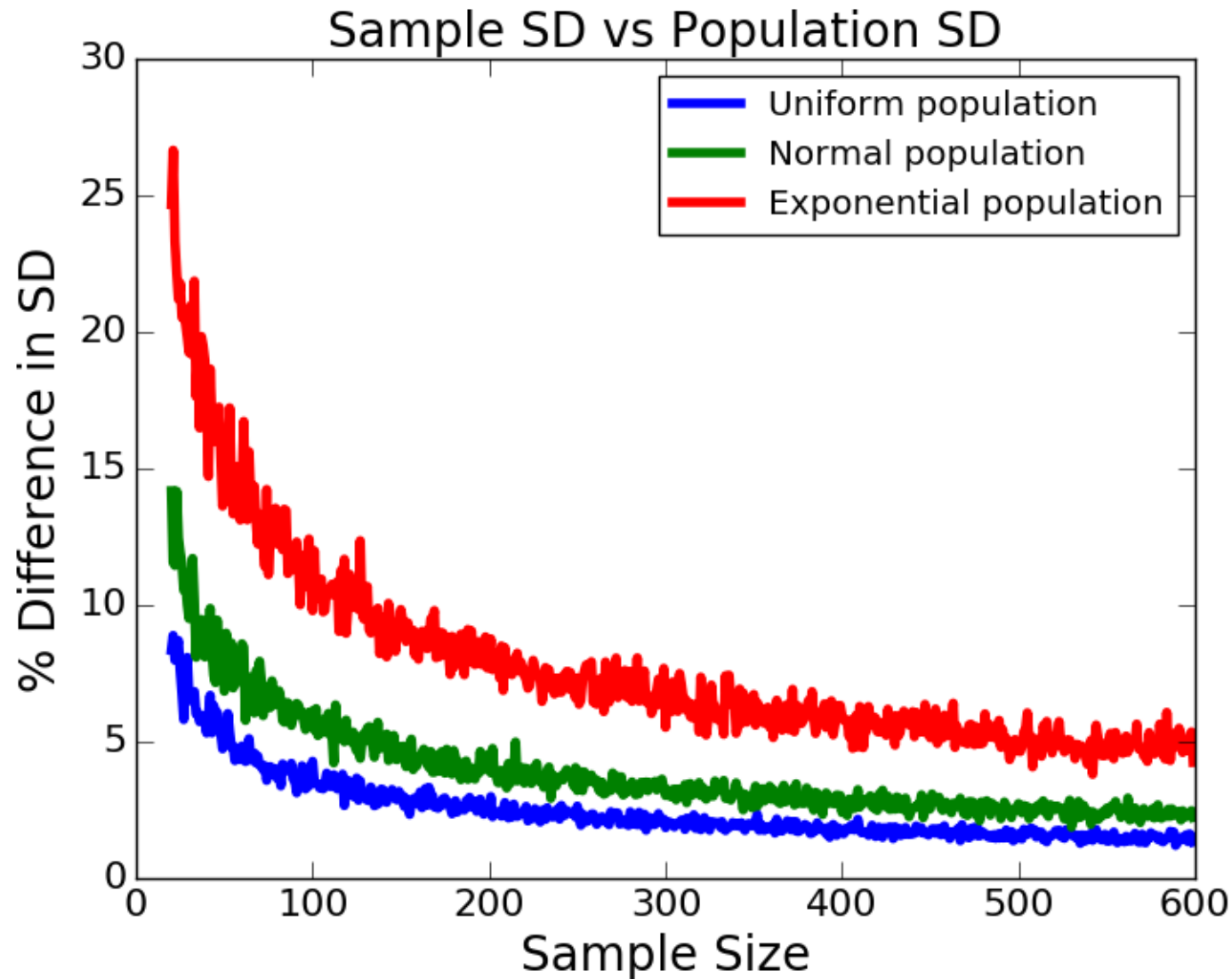
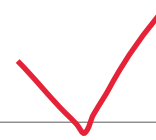
```
def plotDistributions():
    uniform, normal, exp = [], [], []
    for i in range(100000):
        uniform.append(random.random())
        normal.append(random.gauss(0, 1))
        exp.append(random.expovariate(0.5))
    makeHist(uniform, 'Uniform', 'Value', 'Frequency')
    pylab.figure()
    makeHist(normal, 'Gaussian', 'Value', 'Frequency')
    pylab.figure()
    makeHist(exp, 'Exponential', 'Value', 'Frequency')
```

Three Different Distributions

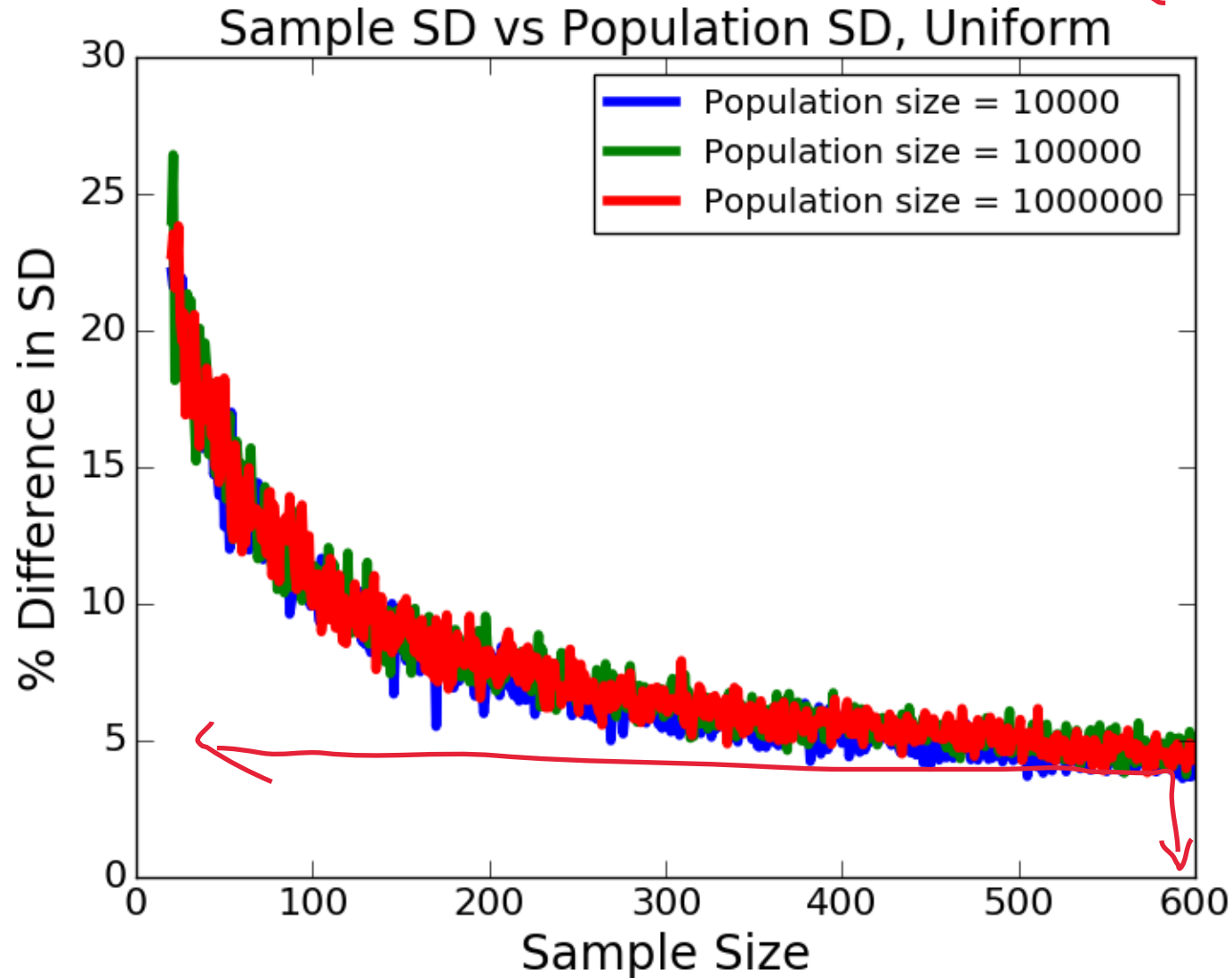


Skew is a measure of the asymmetry of a probability distribution





Does Distribution Matter?



Does Population Size Matter?



To Estimate Mean from a Single Sample

- 1) Choose sample size based on estimate of skew in population 
 - 2) Chose a random sample from the population
 - 3) Compute the mean and standard deviation of that sample
 - 4) Use the standard deviation of that sample to estimate the SE 
 - 5) Use the estimated SE to generate confidence intervals around the sample mean 
- 

Works great when we choose independent random samples.
Not always so easy to do.