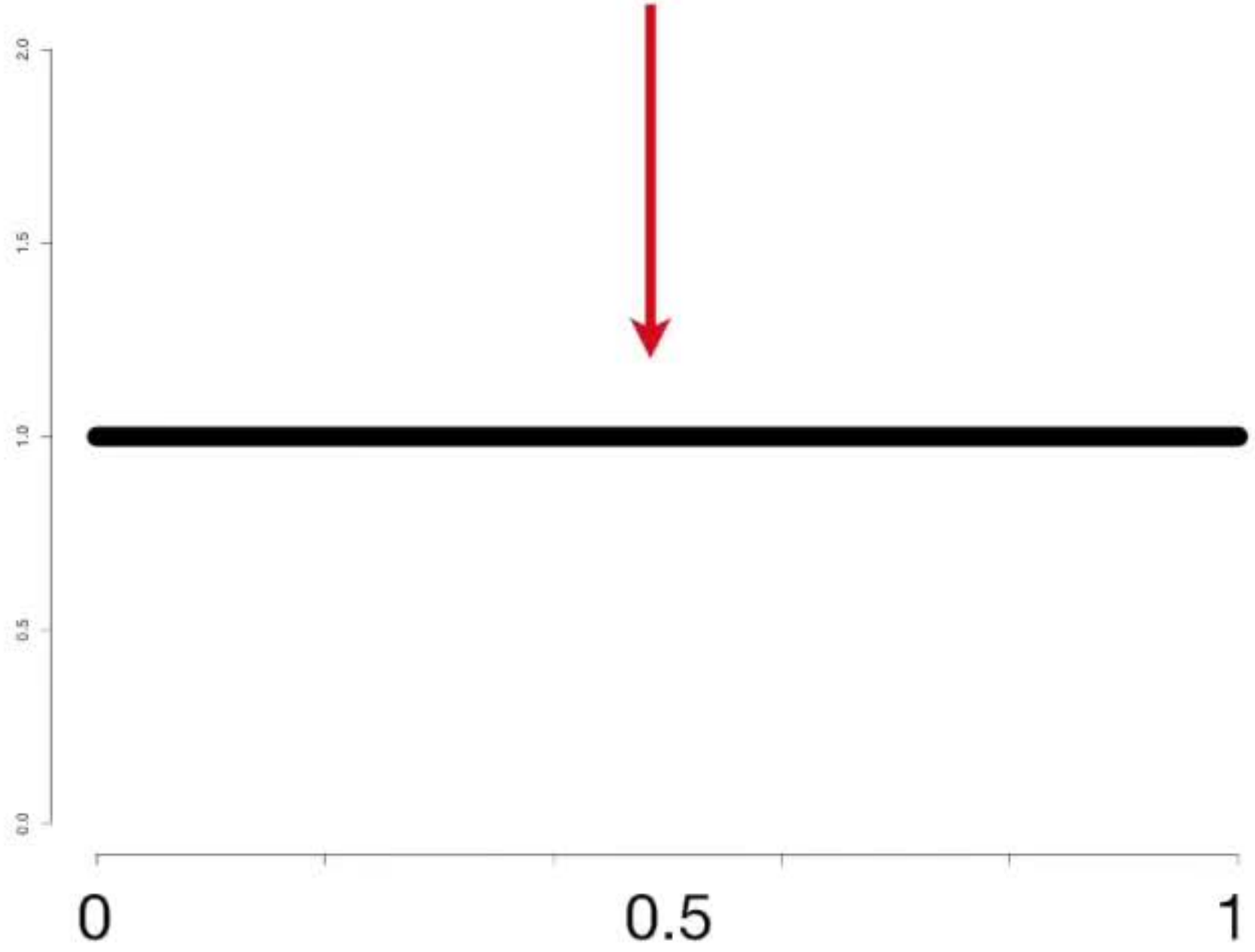
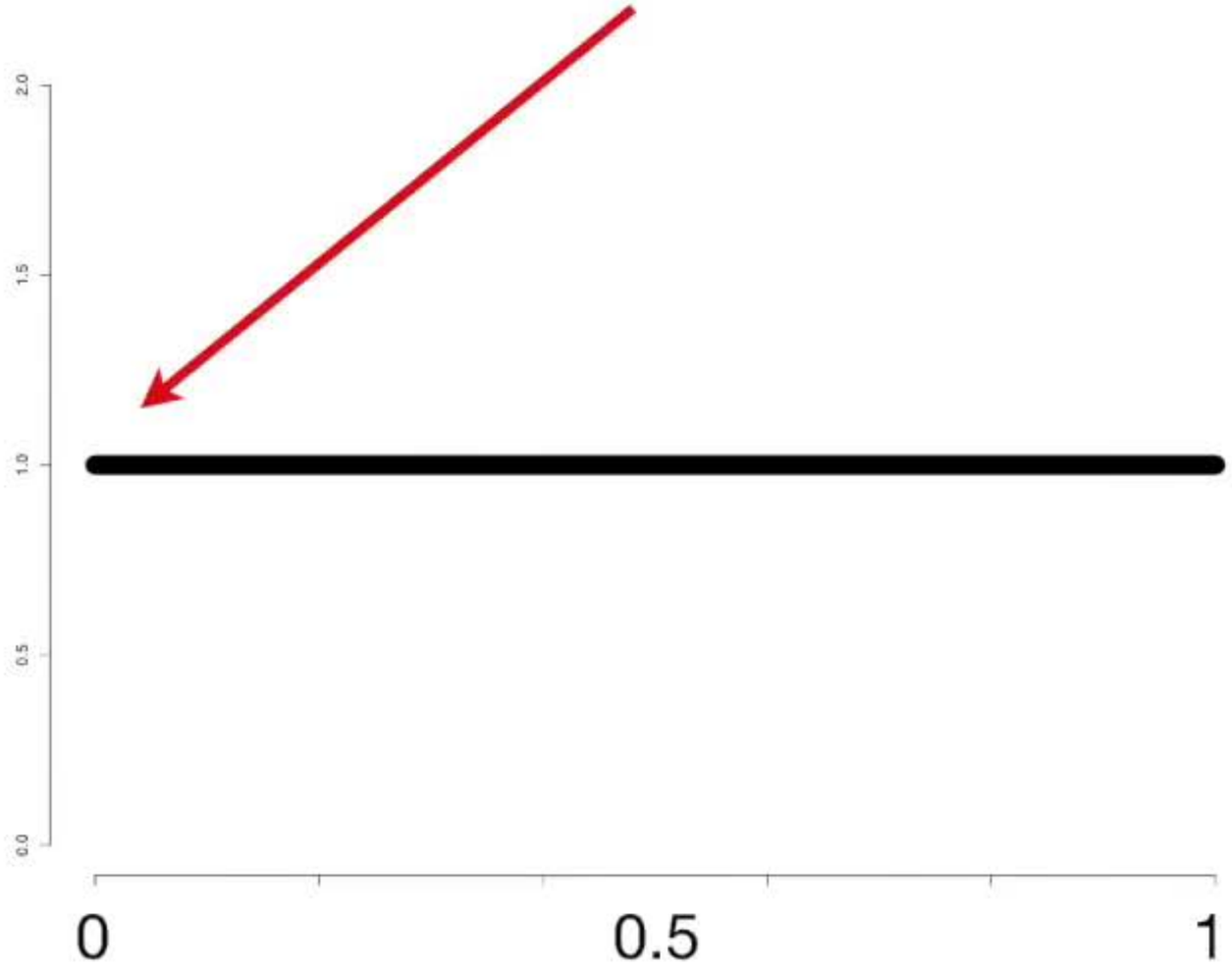
The Central Limit Theorem is the basis for a lot of statistics and the good news is that it is a pretty simple concept.

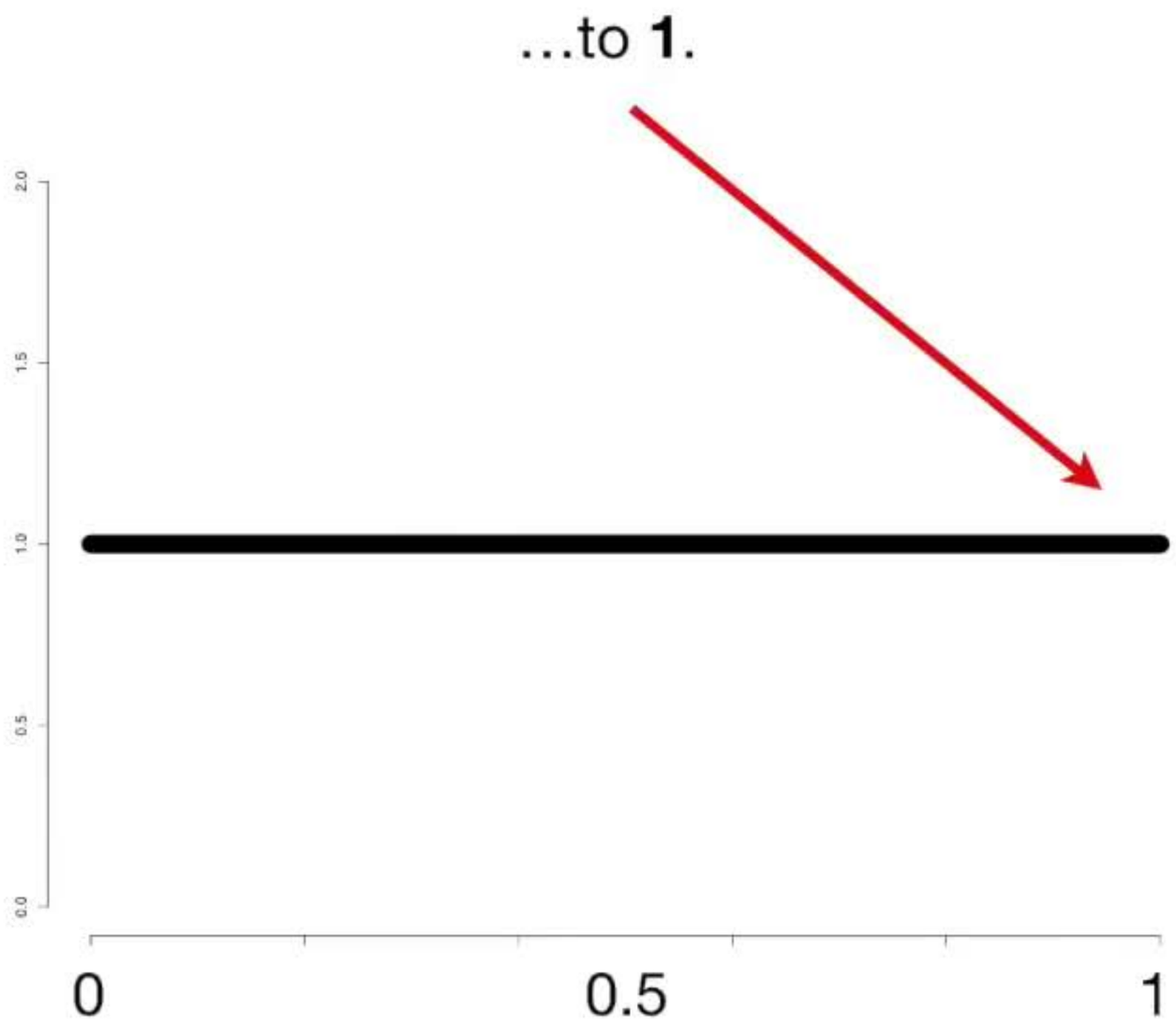
Like most things in statistics, I think The Central Limit Theorem is easiest to understand if we look at some examples.

So let's start with a Uniform Distribution.

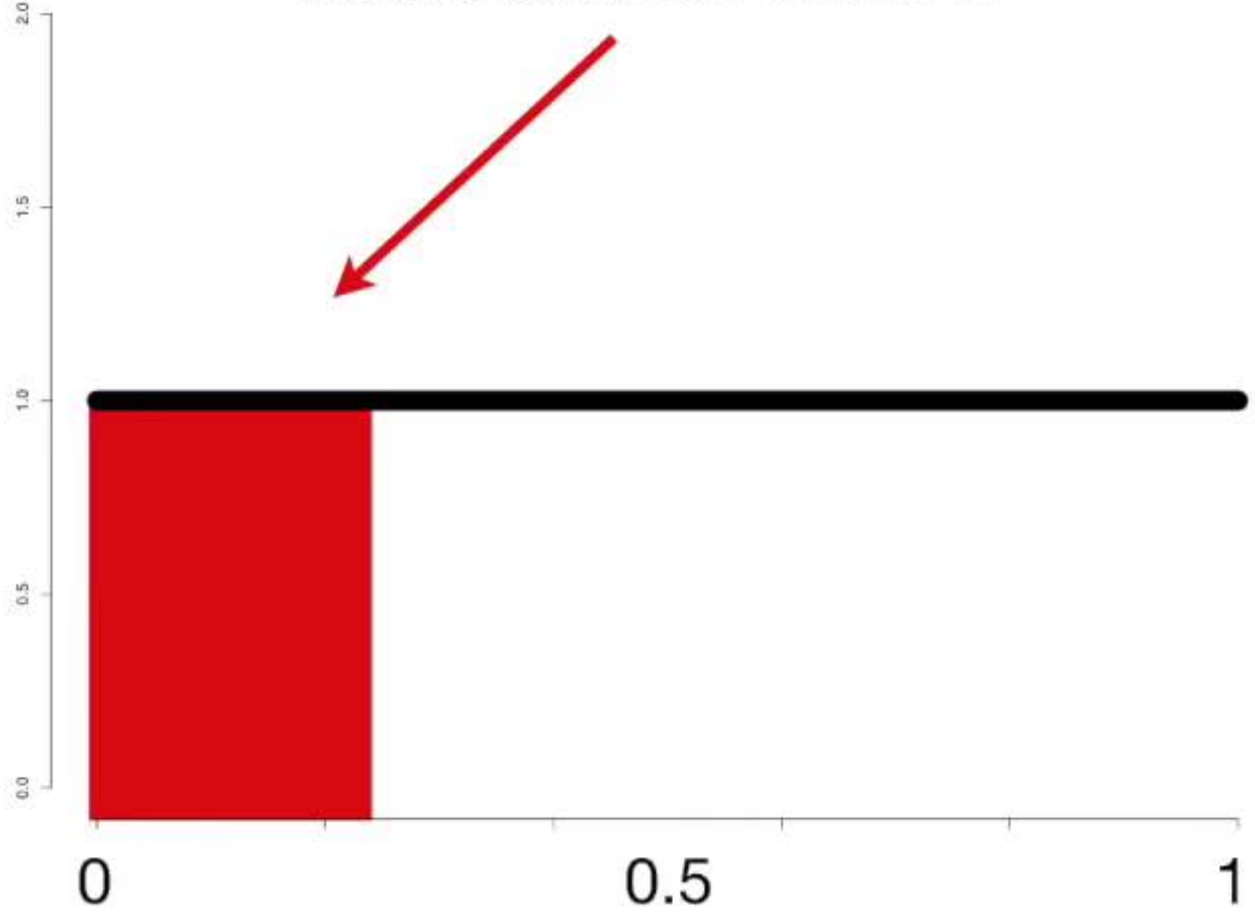


This one goes from **0**…

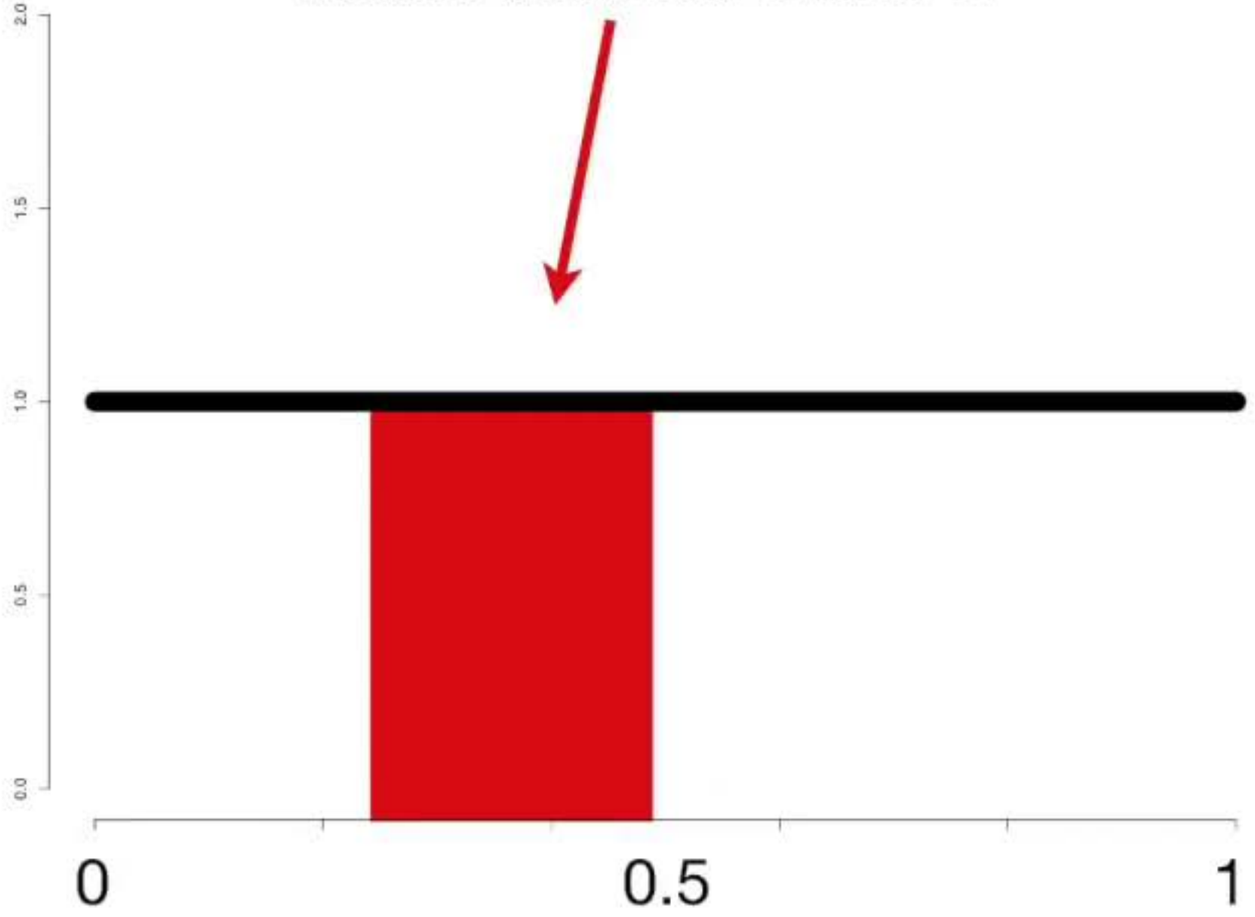



…to **1**.
2.0
1.5
1.0
0.5
0.0
0
0.5
1

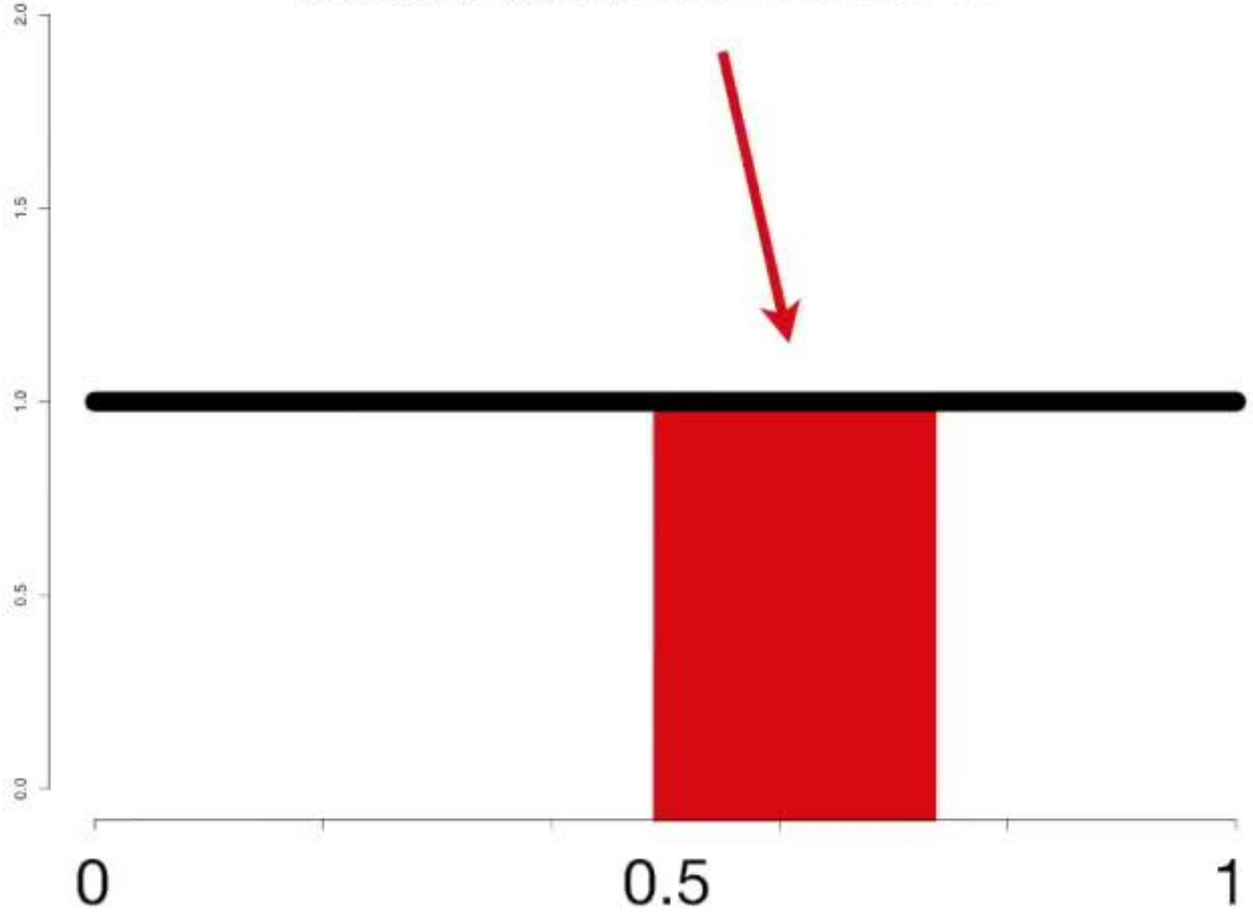
It's called the uniform distribution because there is an equal probability of selecting values between 0 and 1.



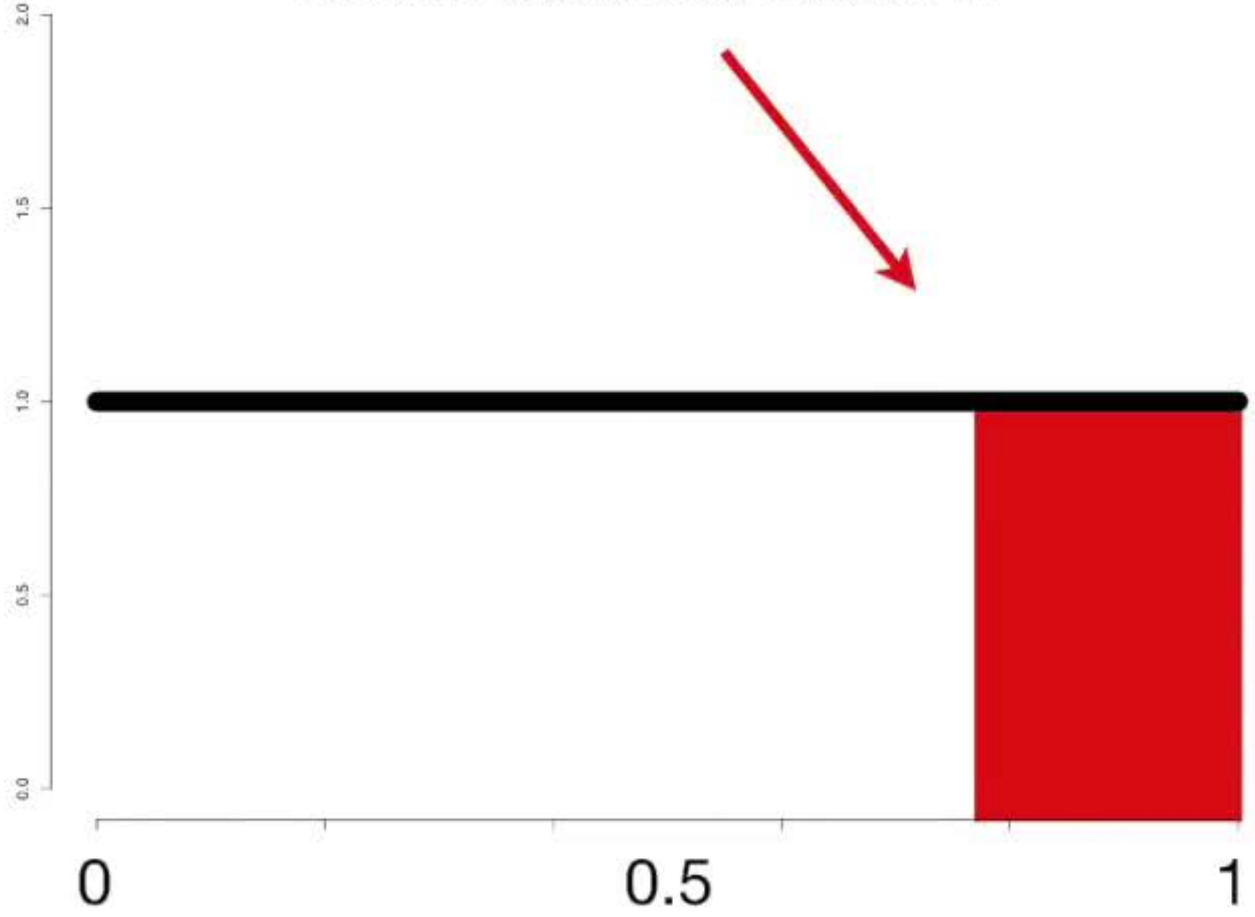
It's called the uniform distribution because there is an equal probability of selecting values between 0 and 1.



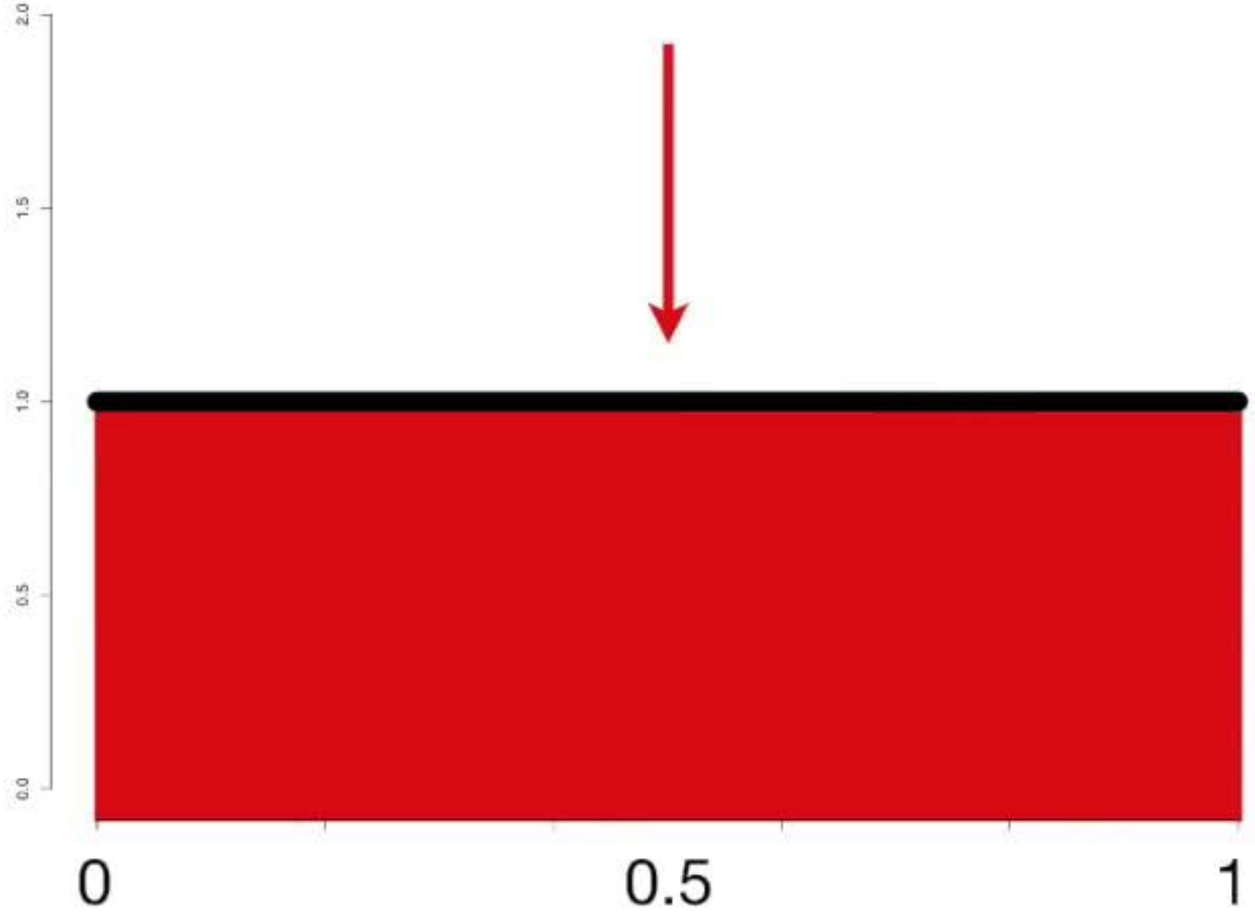
It's called the uniform distribution because there is an equal probability of selecting values between 0 and 1.

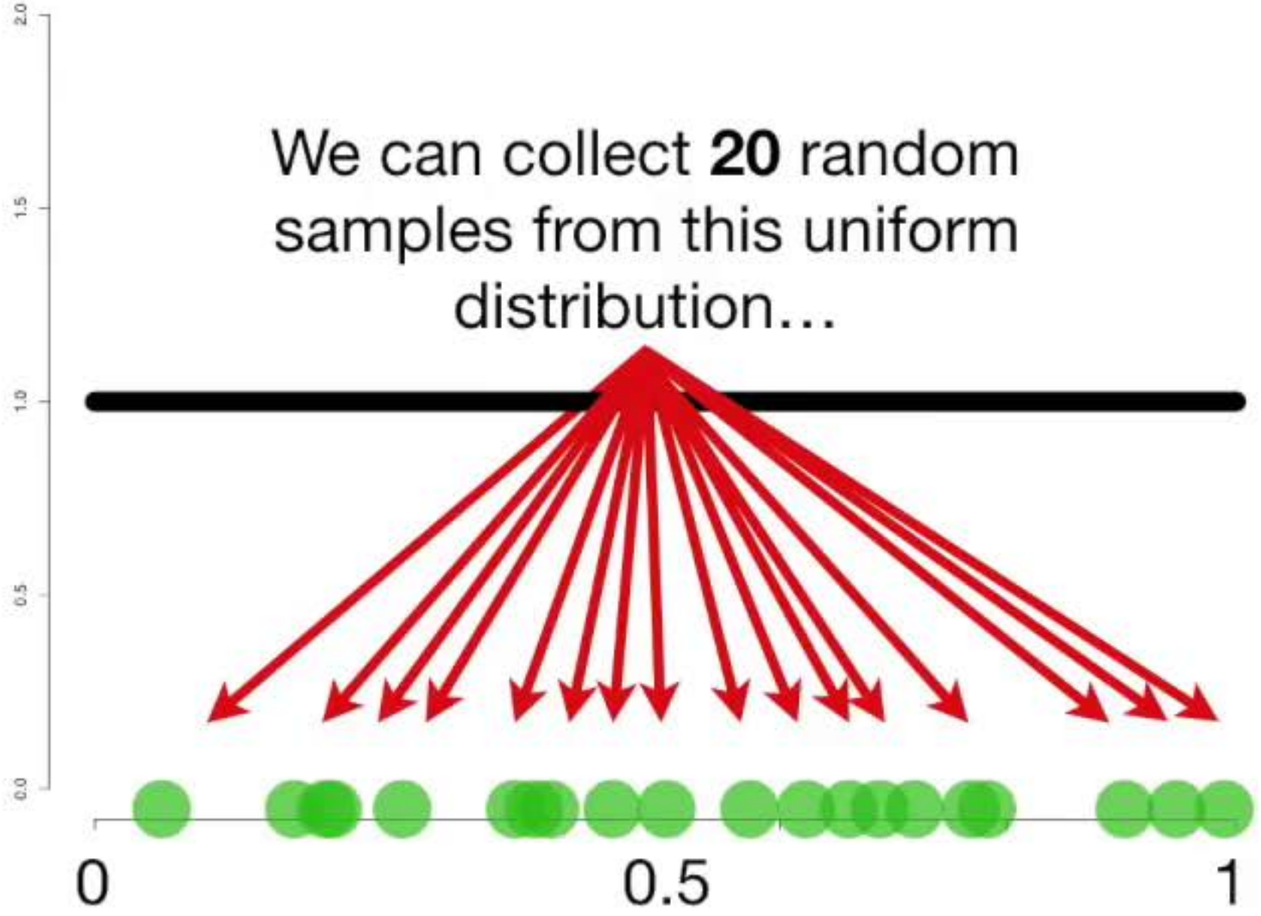


It's called the uniform distribution because there is an equal probability of selecting values between 0 and 1.

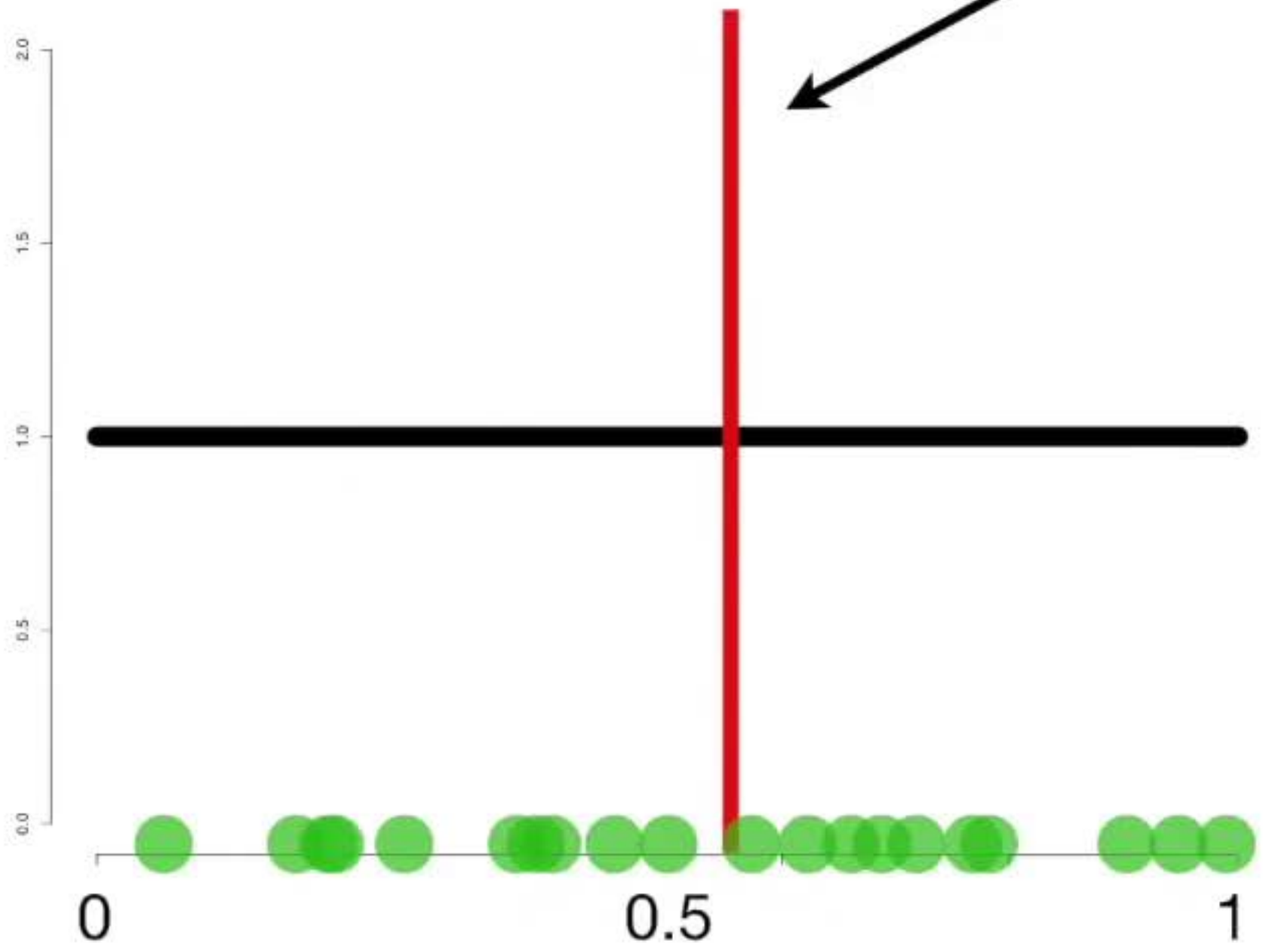


The probabilities are all equal,
and thus, are “uniform”.

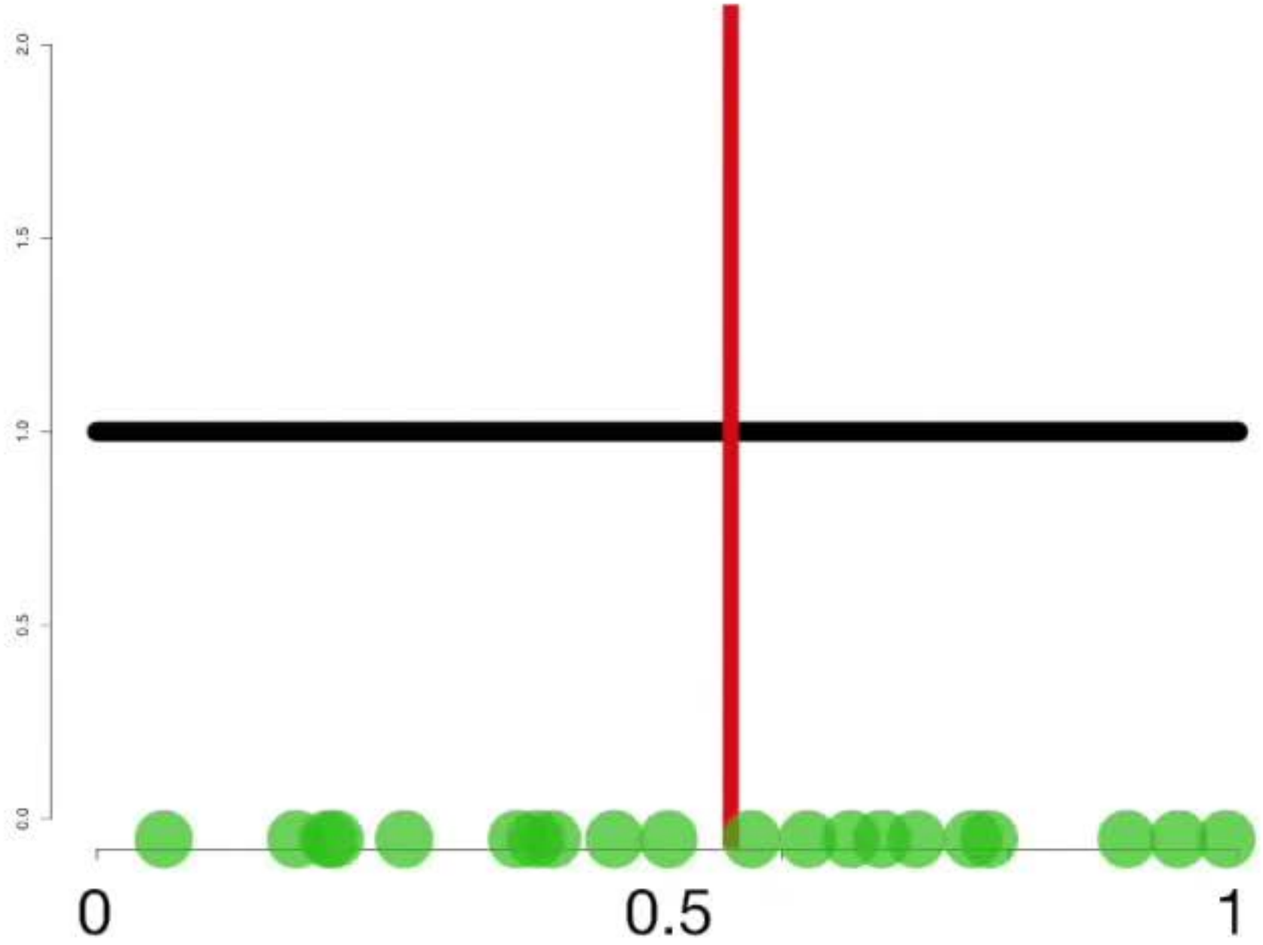


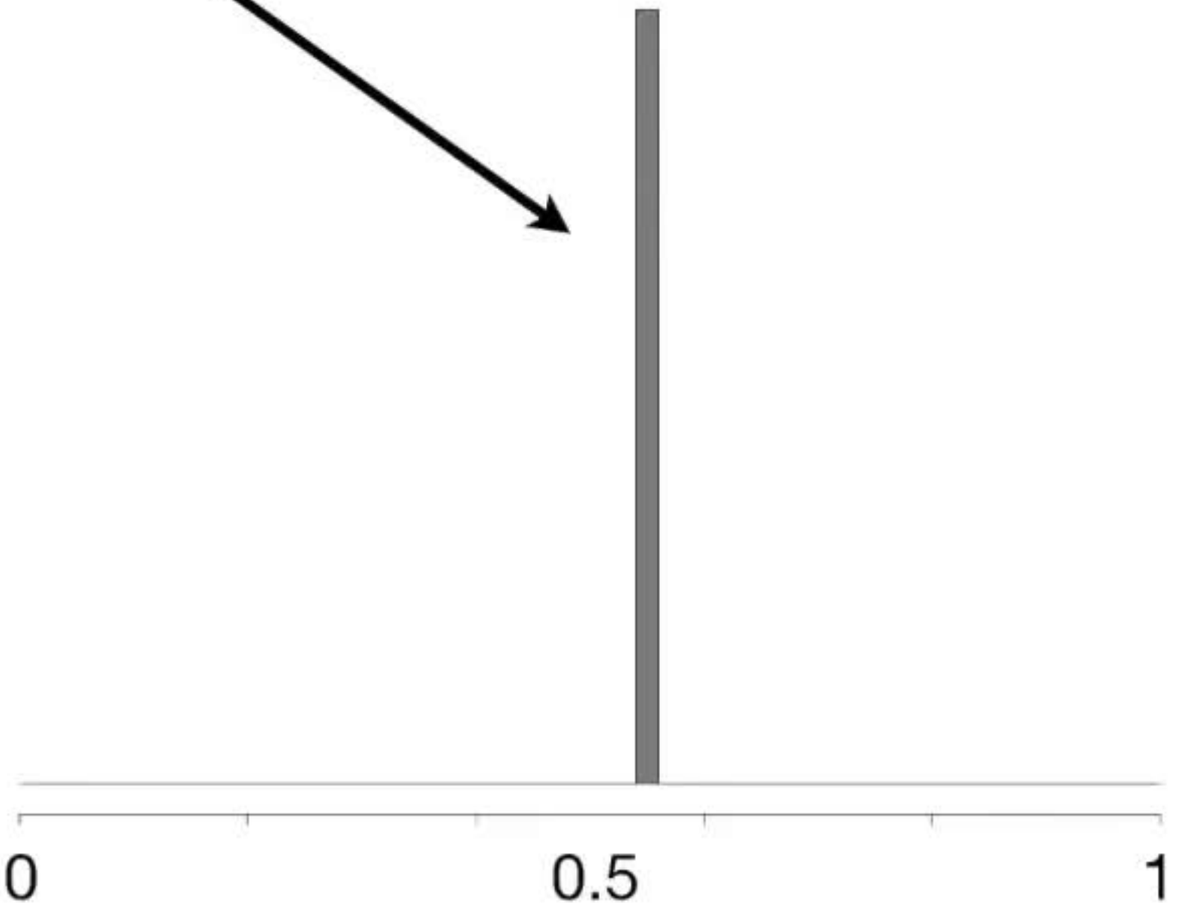

We can collect **20** random samples from this uniform distribution…
2.0
1.5
1.0
0.5
0.0
0
0.5
1

…and then calculate the mean of the samples.

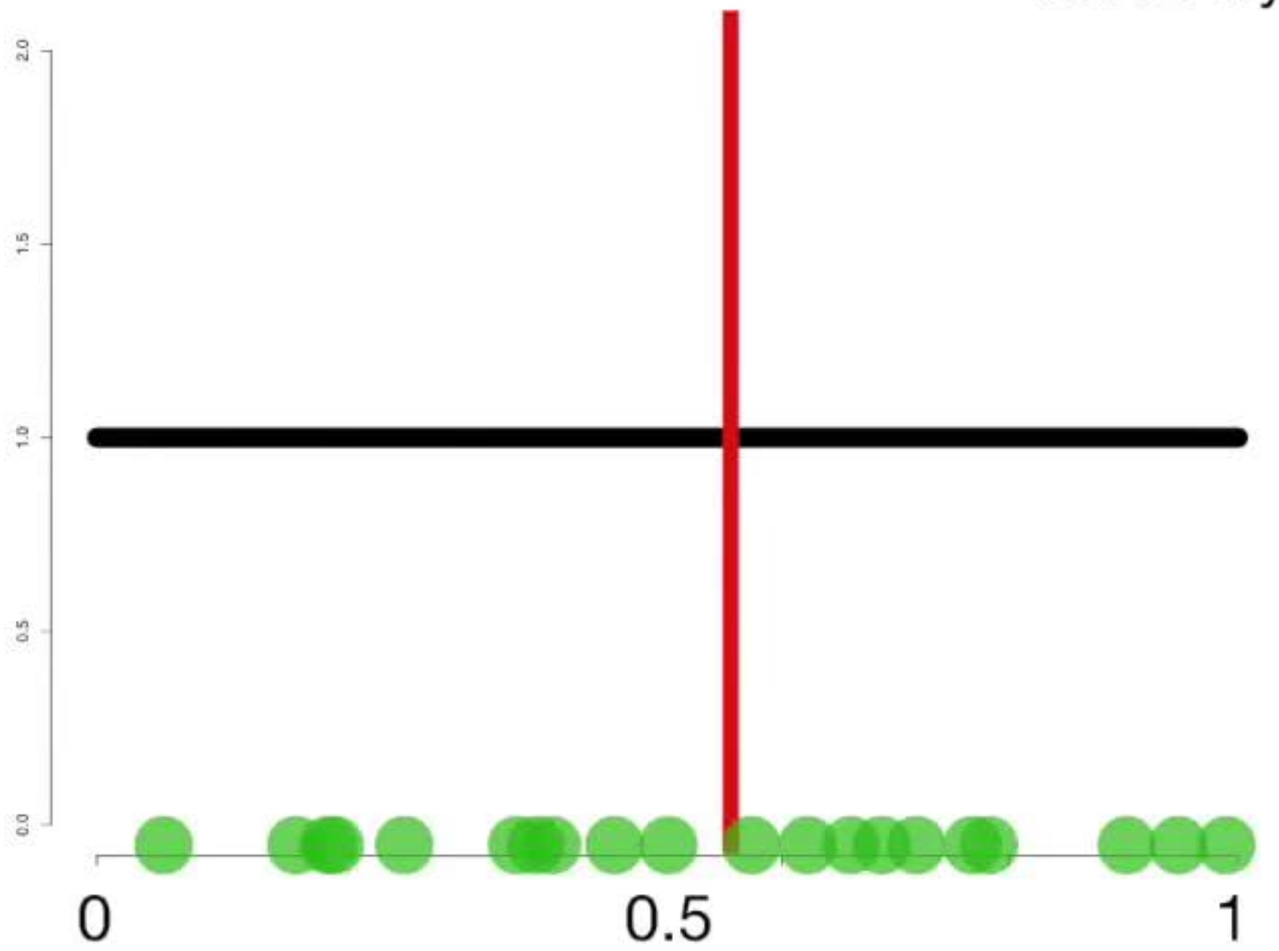


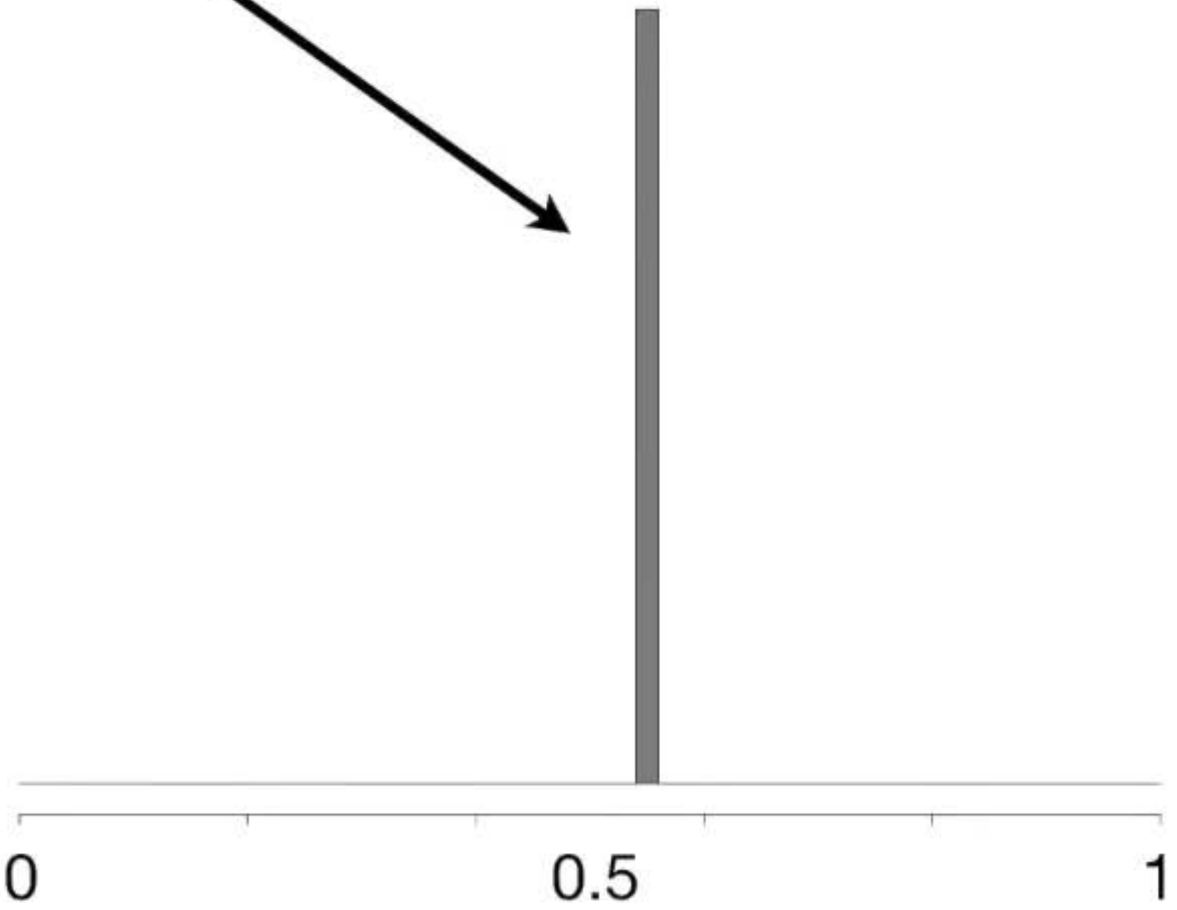
On the right, we can draw a histogram of the mean value.



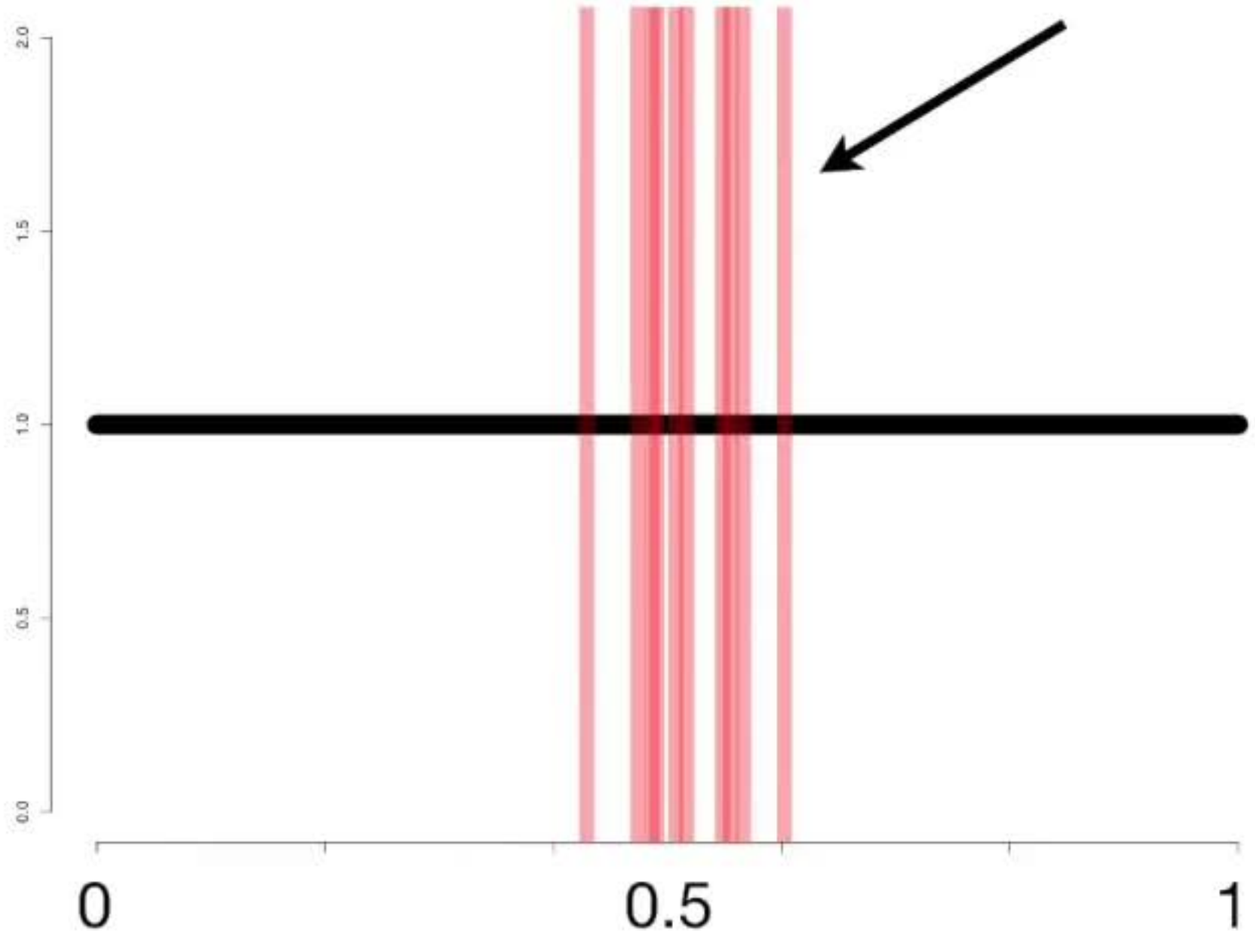


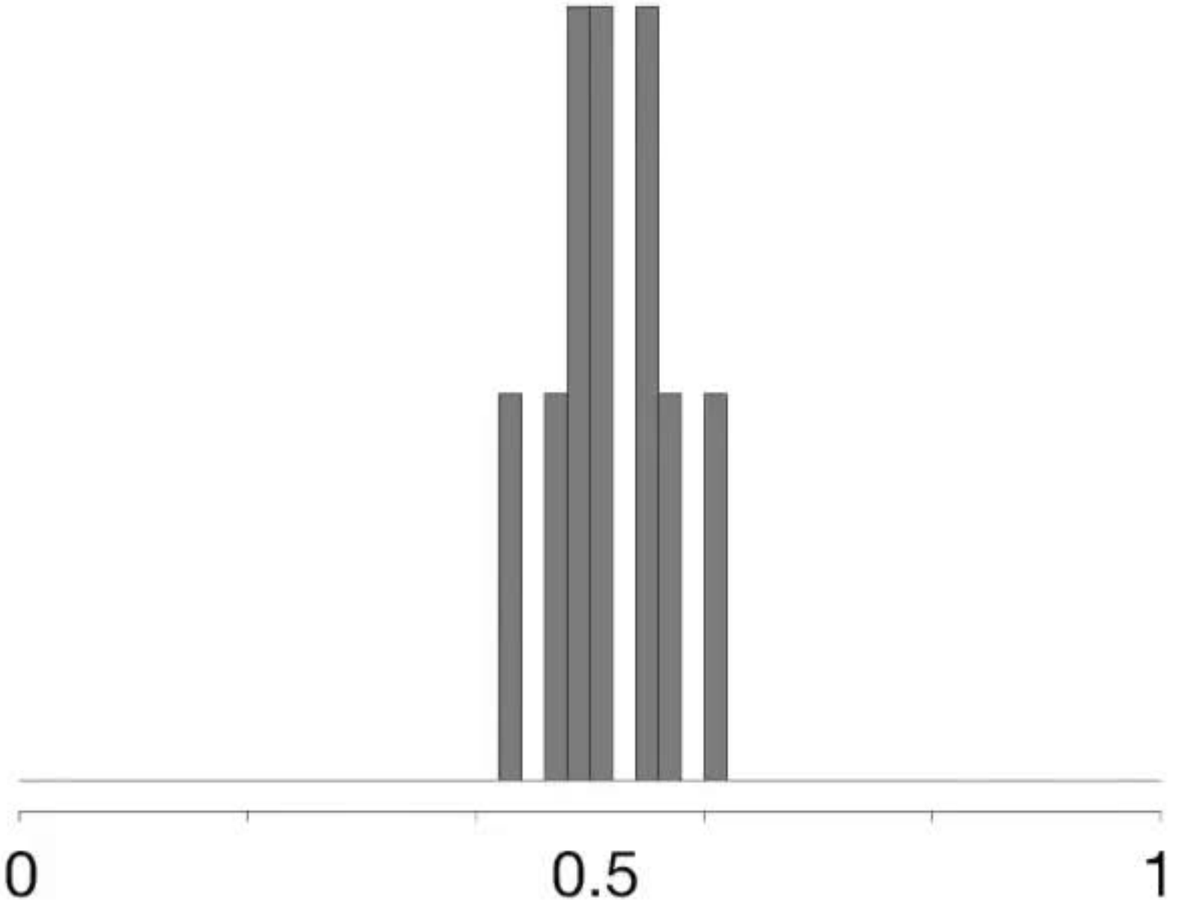
Since we only have one mean value, the histogram isn't very interesting…



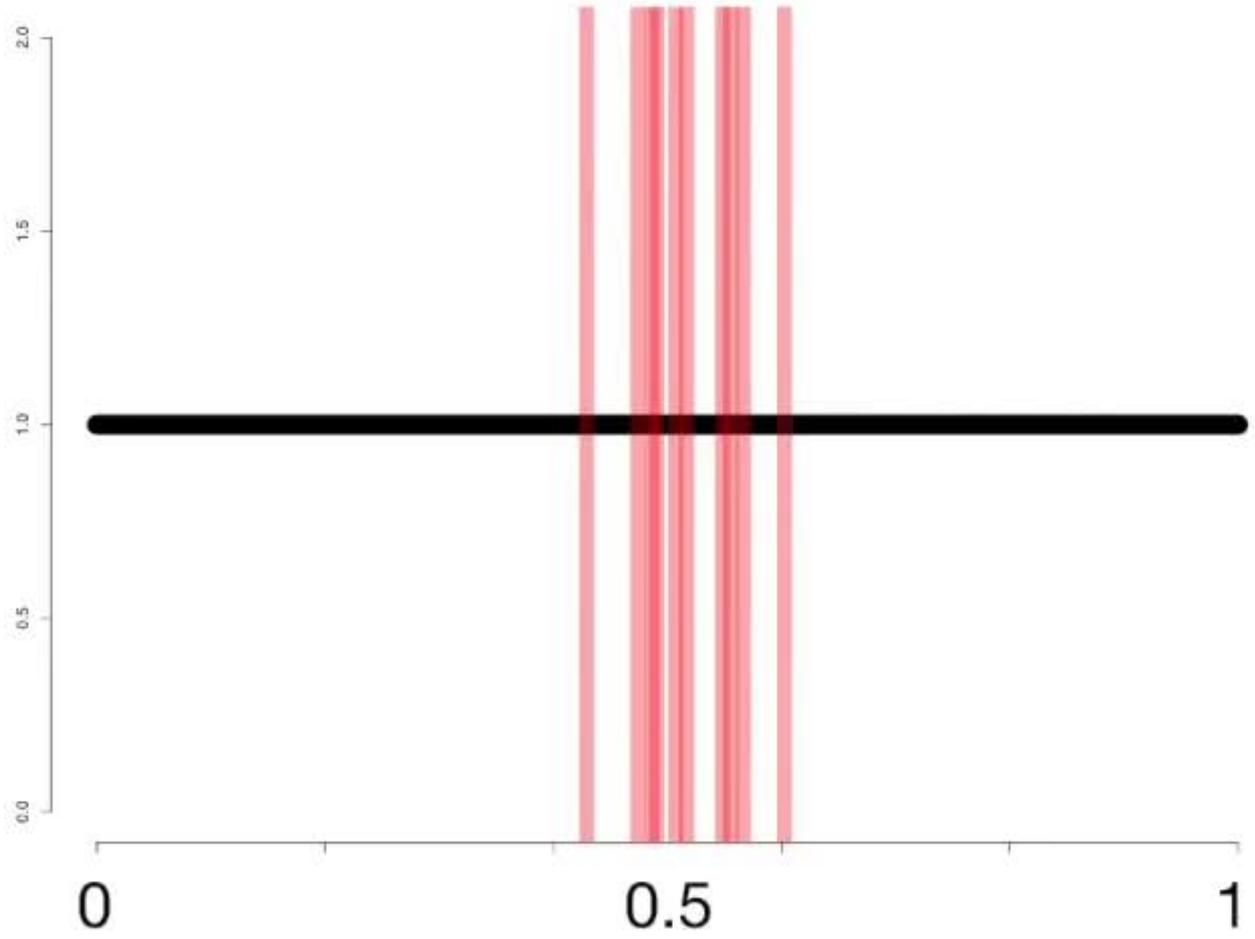


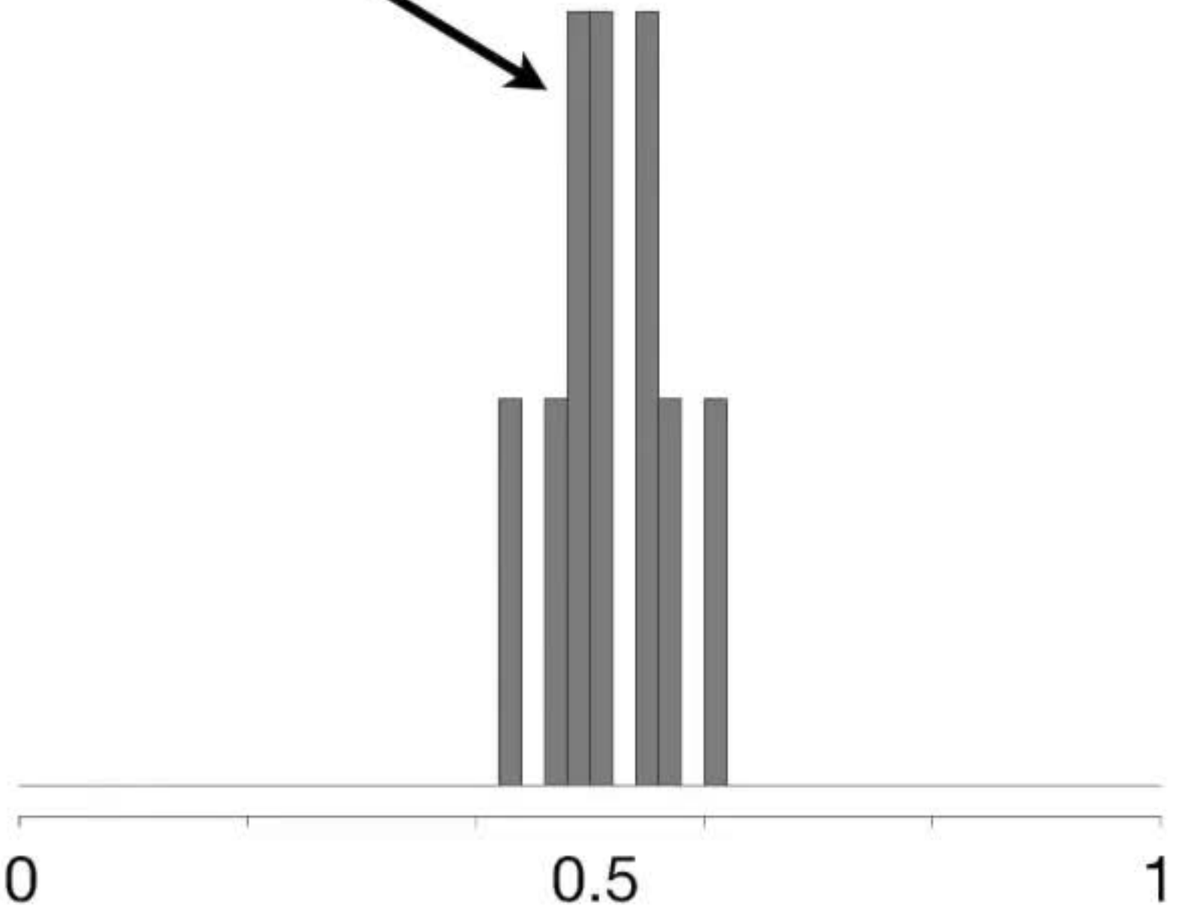
…but after we collect 10 more samples and calculate 10 more means…



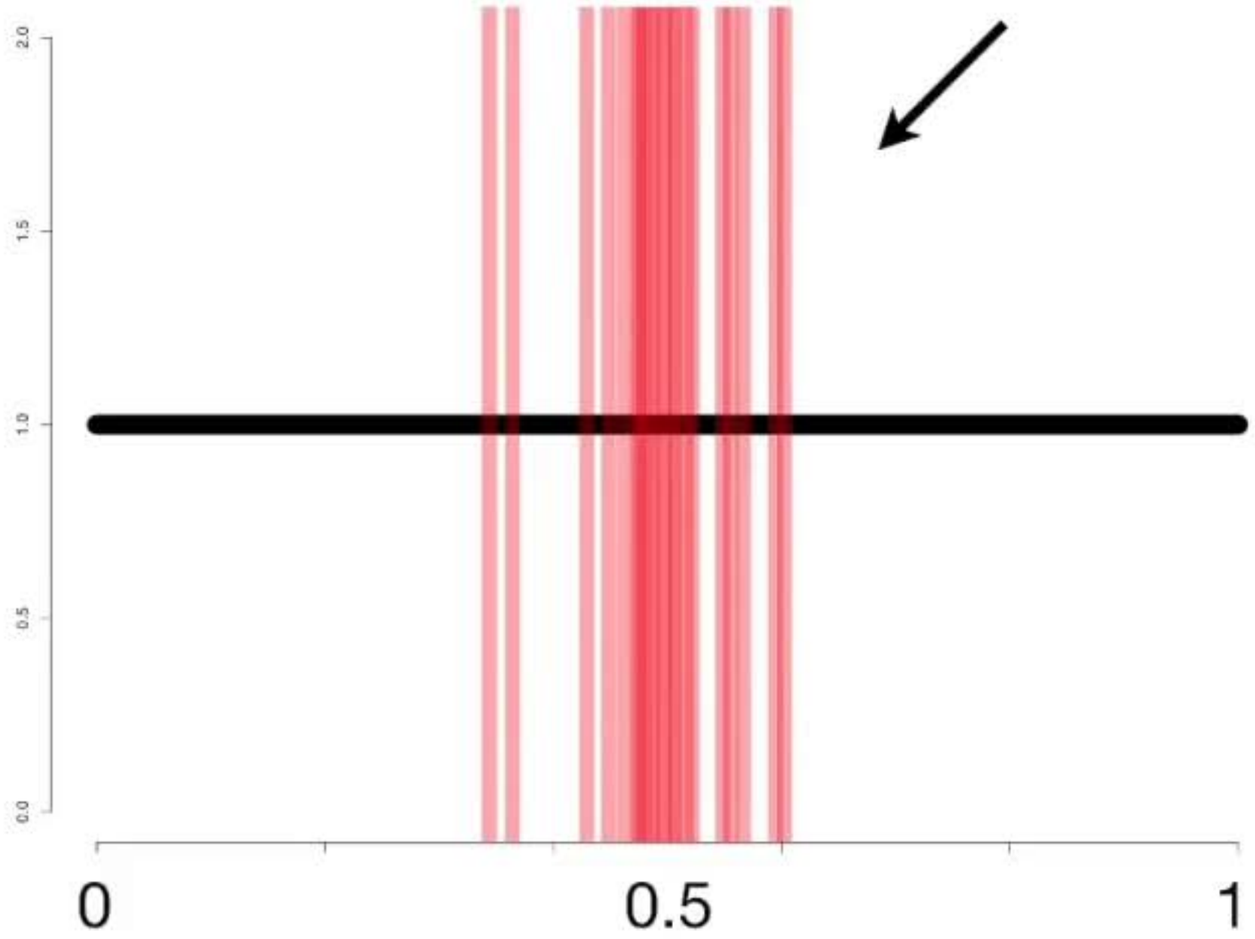


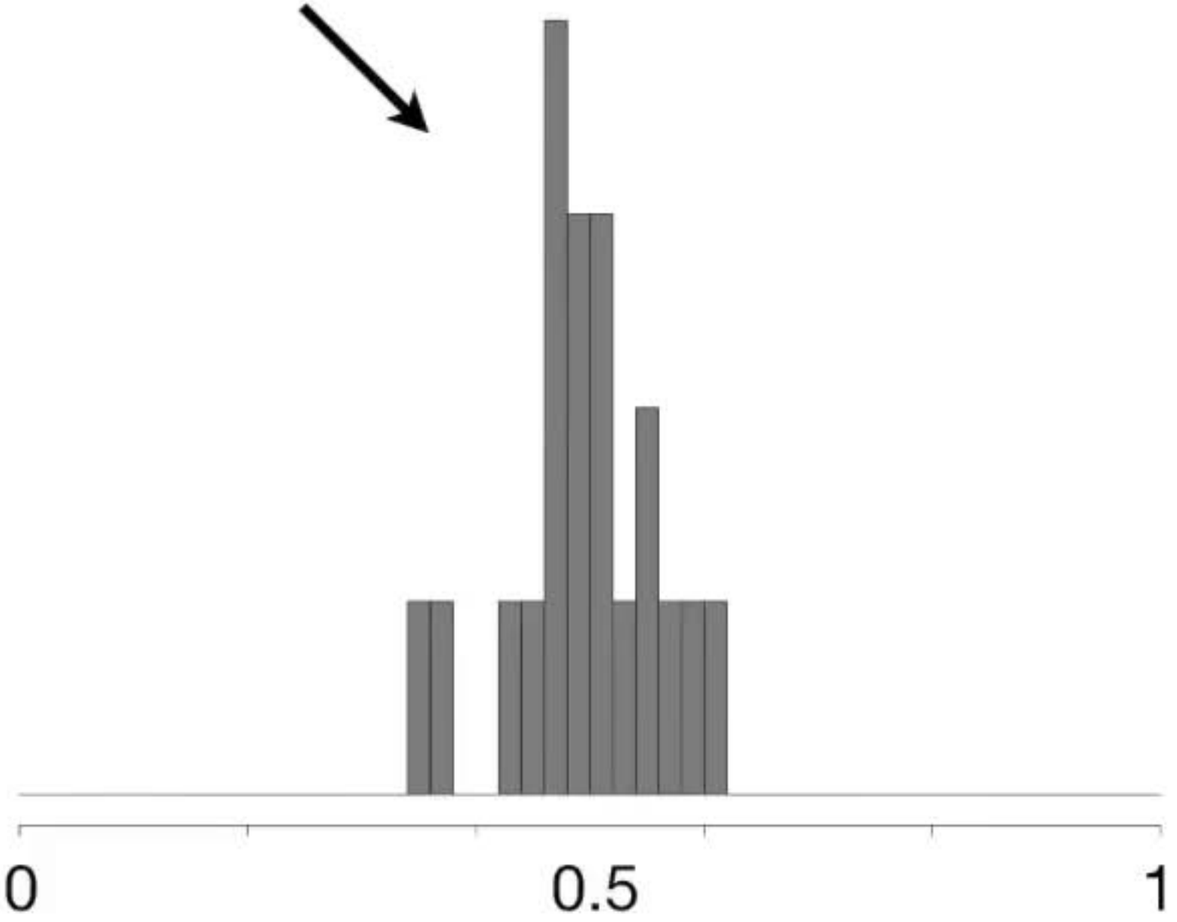
…the histogram starts to look at little more interesting.

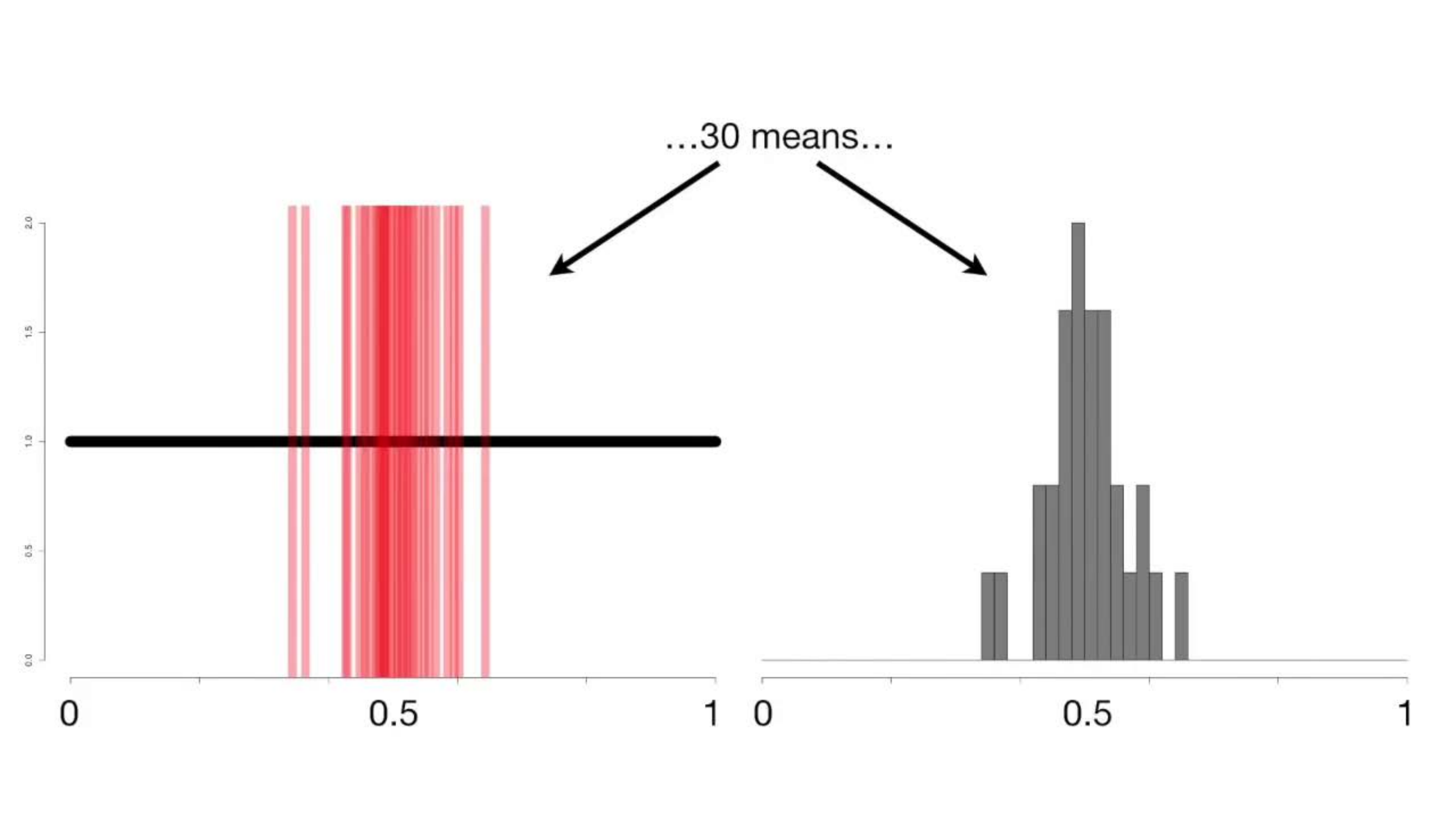


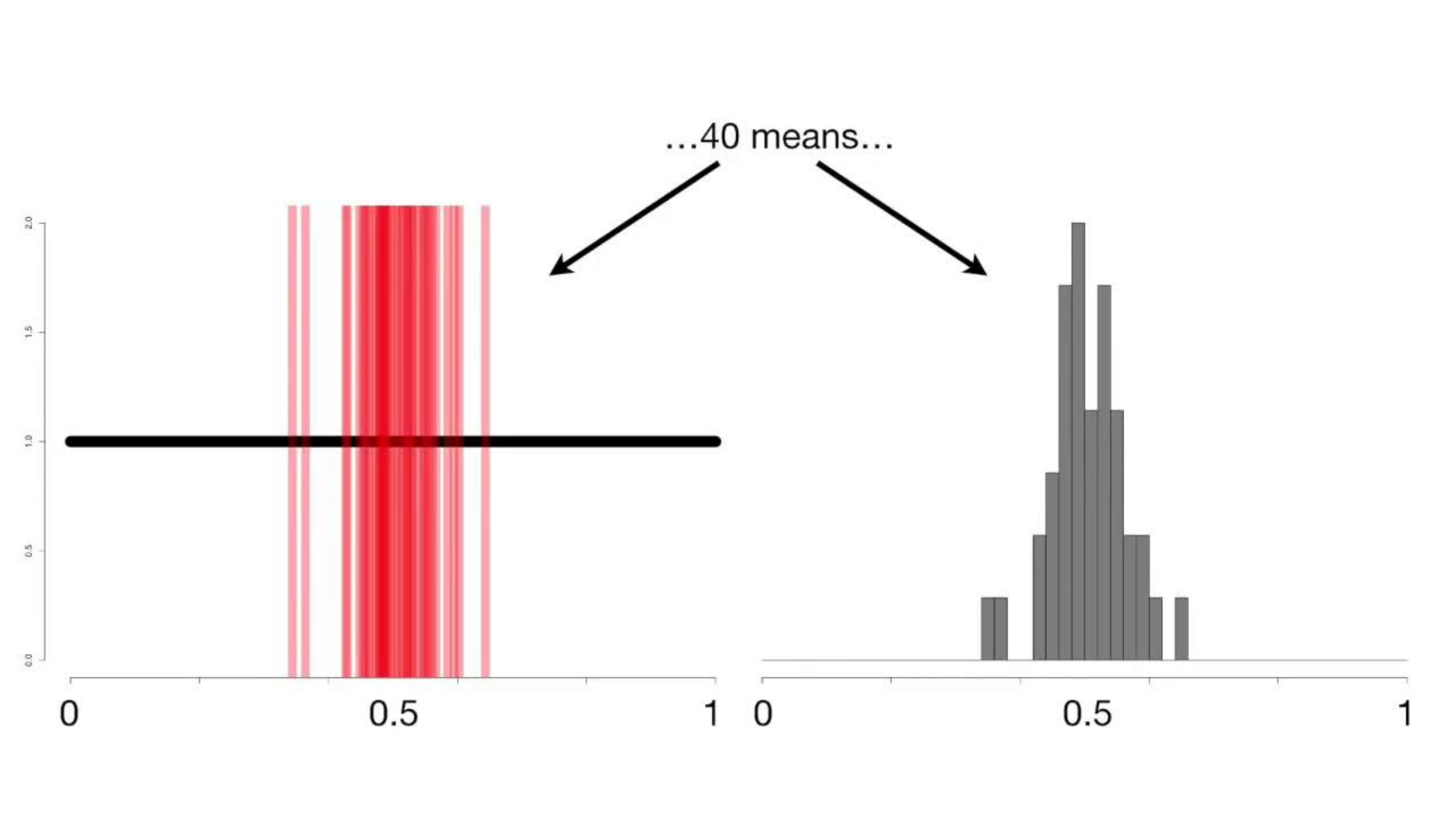


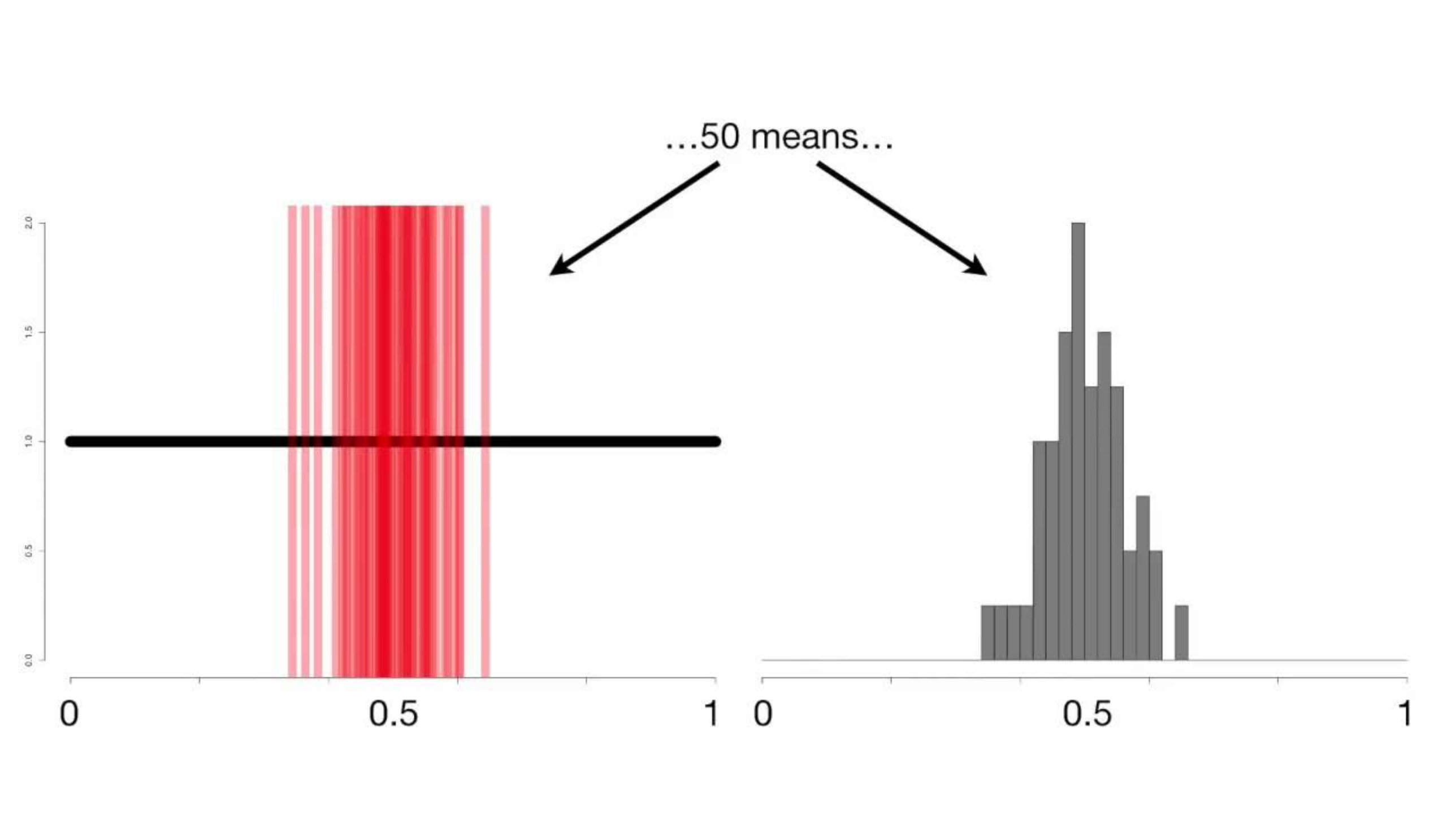
Here's the histogram after collecting 20 samples and calculating 20 means…

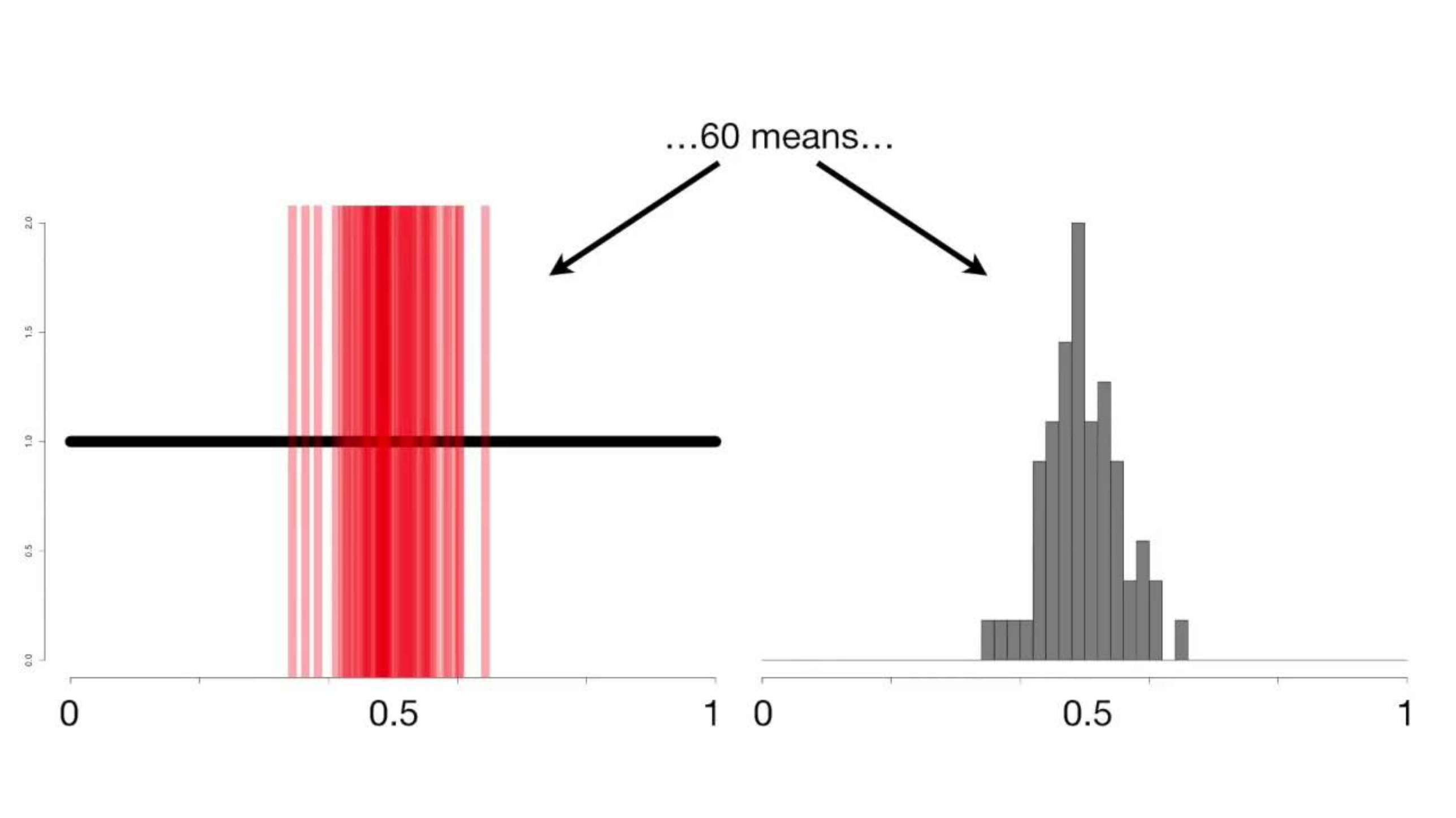


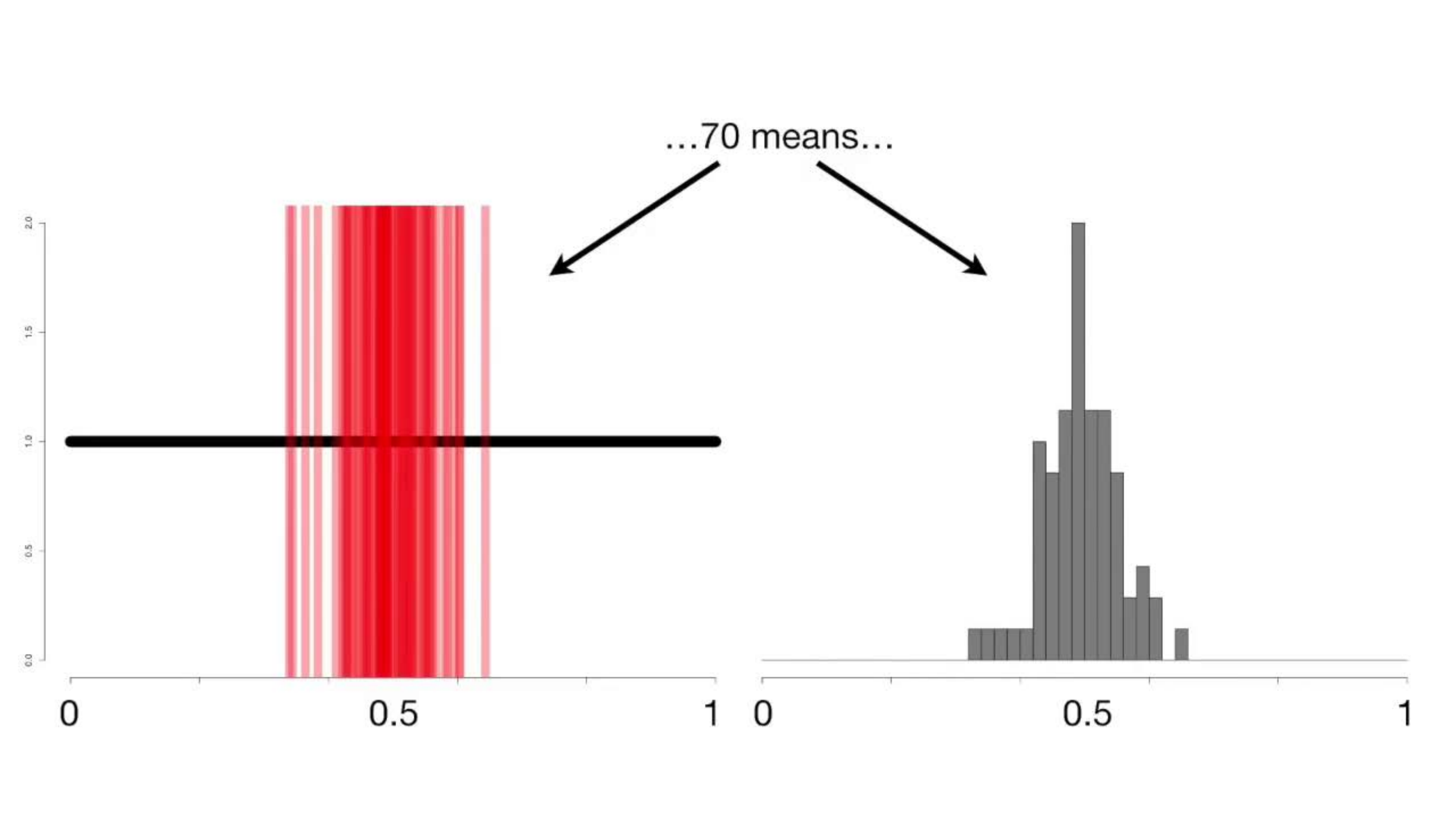


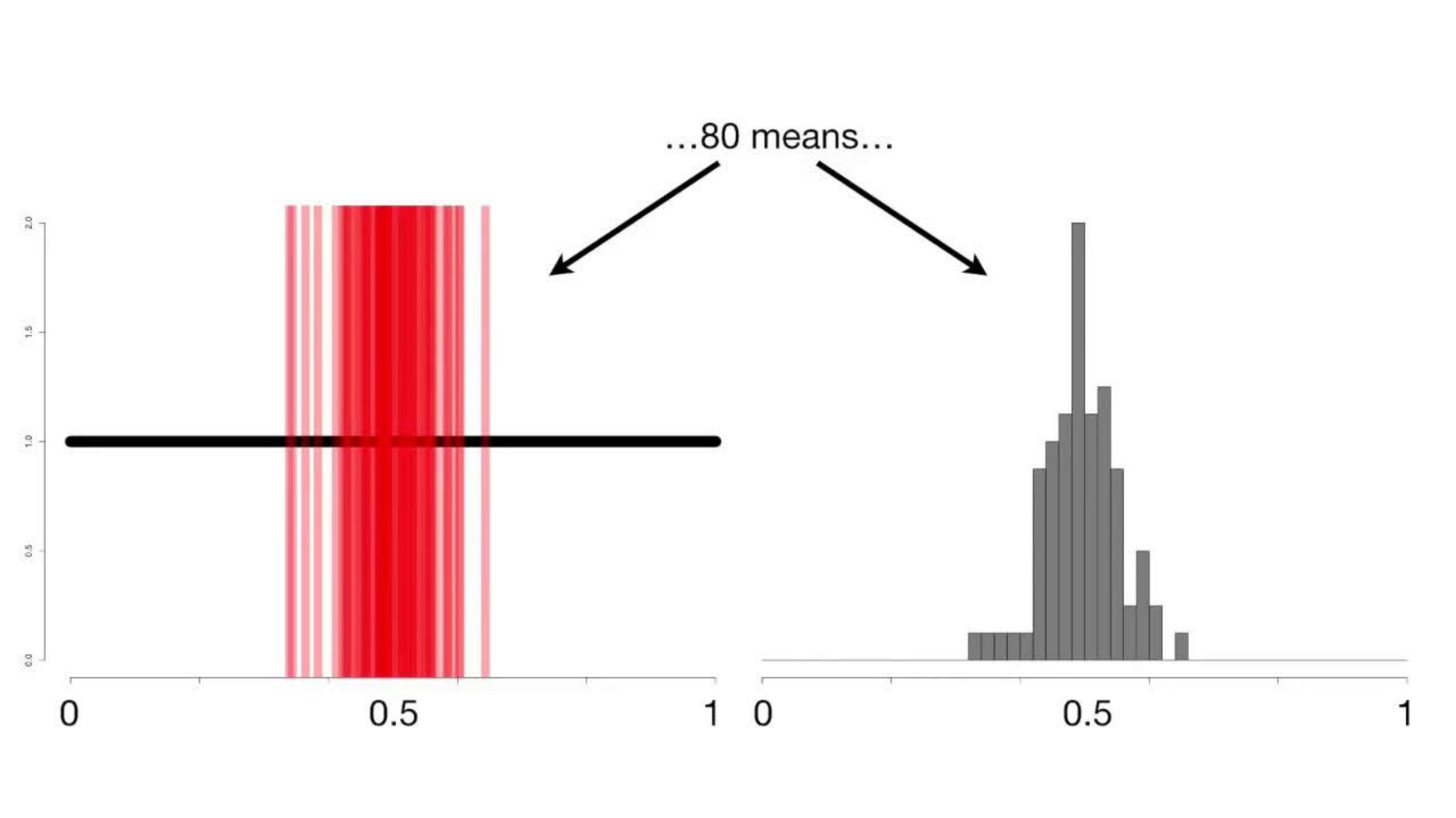

…30 means…
0.0
0.5
1.0
1.5
2.0
0
0.5
1
0
0.5
1

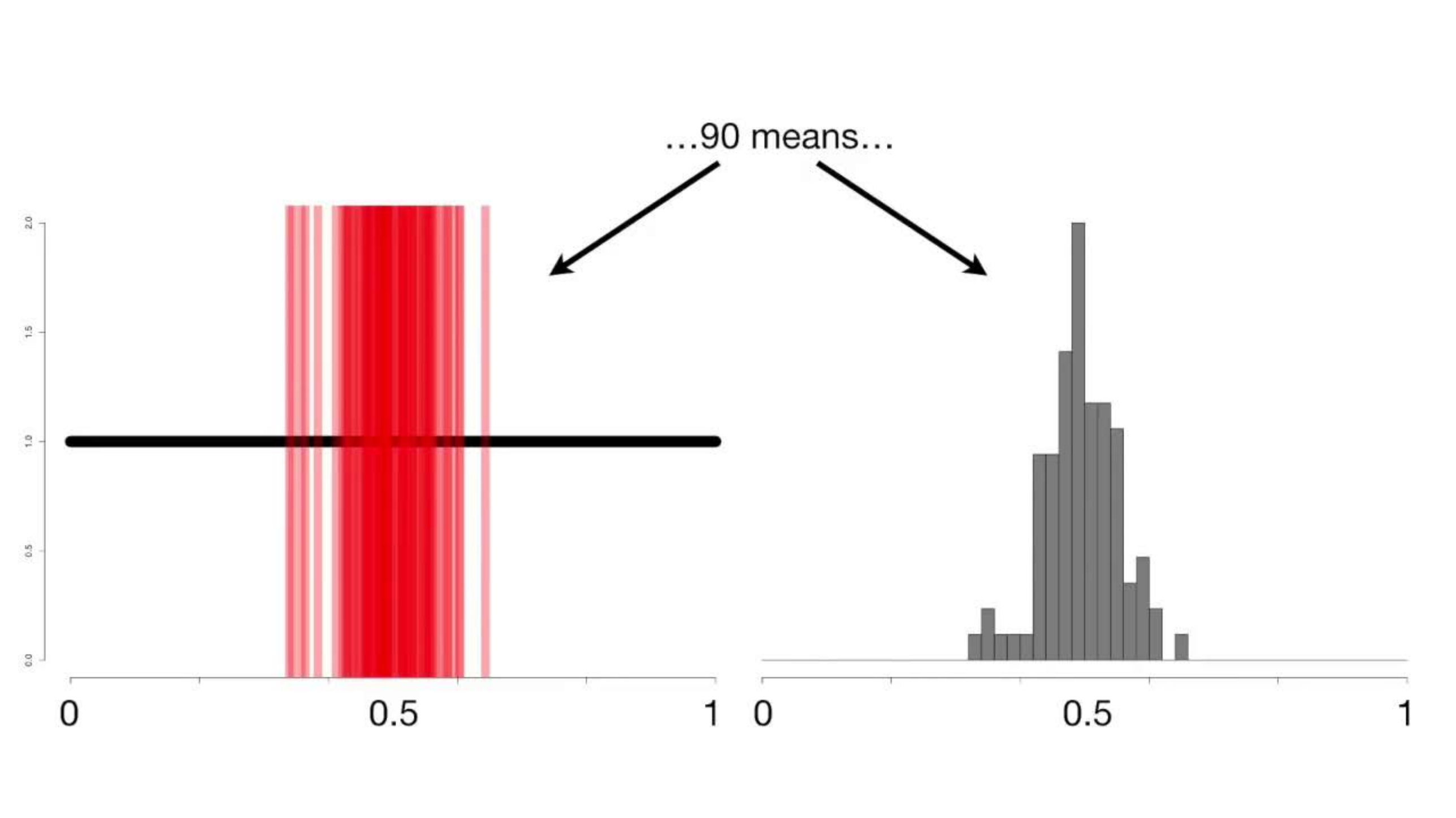

…40 means…
2.0
1.5
1.0
0.5
0.0
0
0.5
1
0
0.5
1

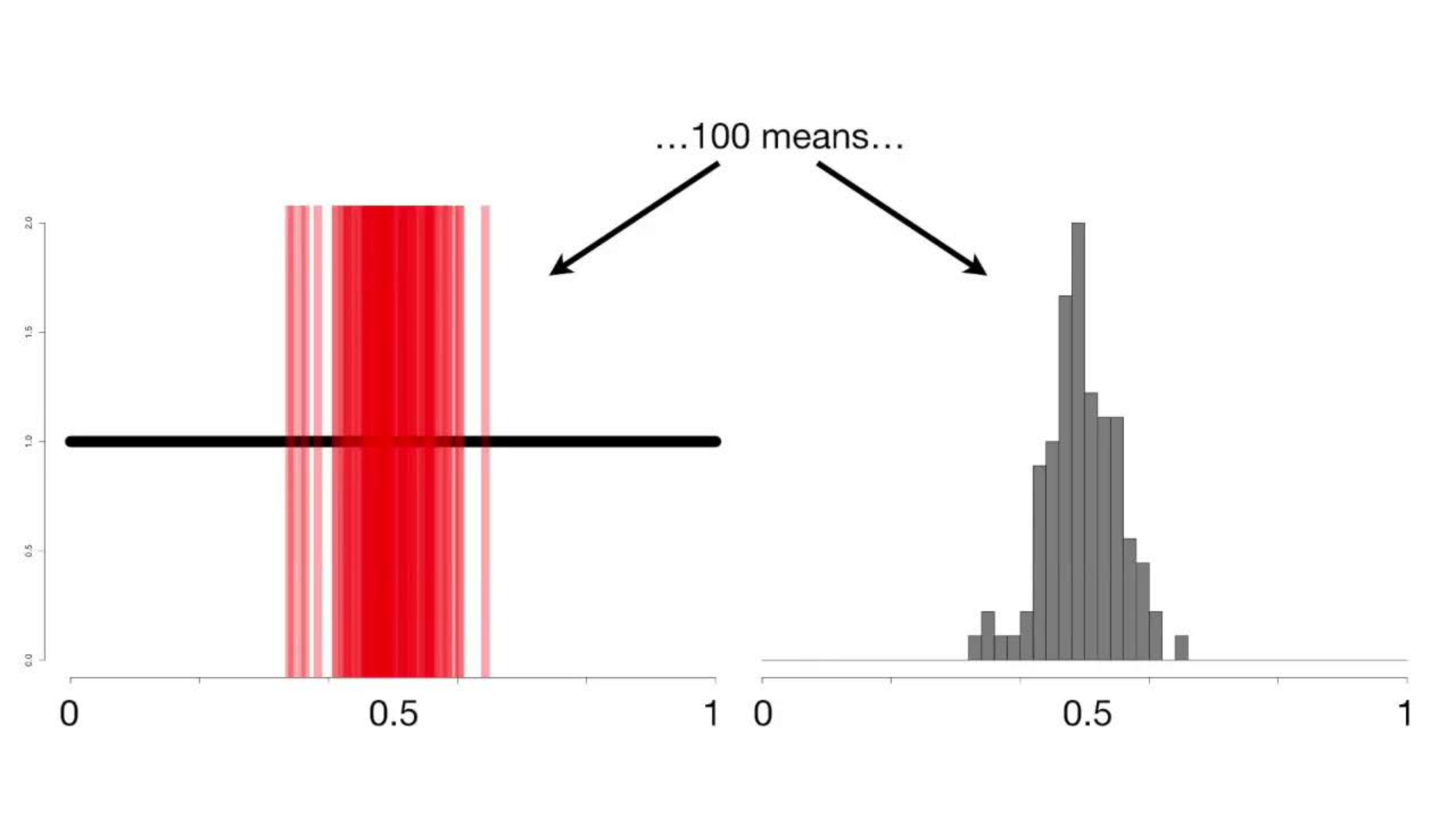

…50 means…
0.0
0.5
1.0
1.5
2.0
0
0.5
1
0
0.5
1


…60 means…
0.0
0.5
1.0
1.5
2.0
0
0.5
1
0
0.5
1

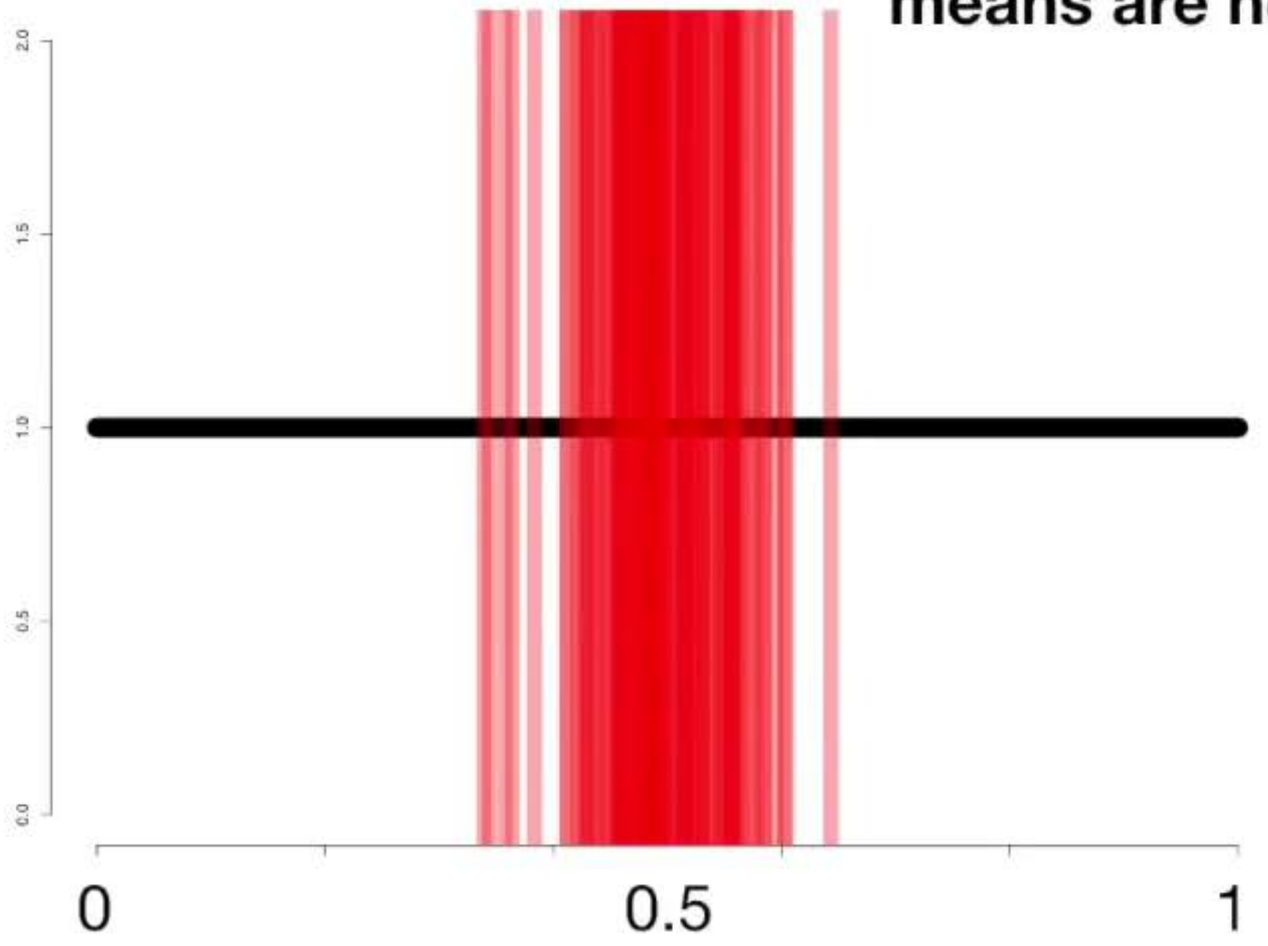

…70 means…
0.0
0.5
1.0
1.5
2.0
0
0.5
1
0
0.5
1

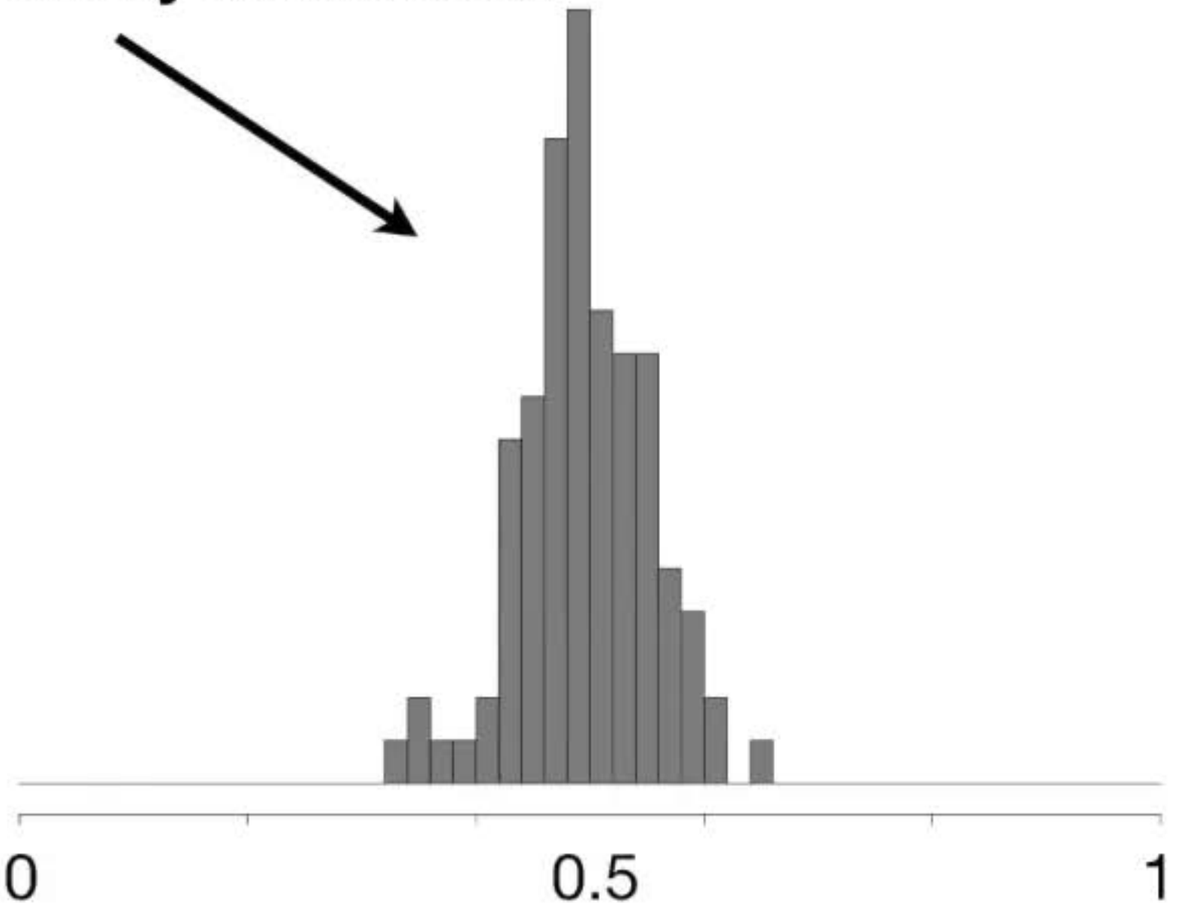

…80 means…
0.0
0.5
1.0
1.5
2.0
0
0.5
1
0
0.5
1


…90 means…
0.0
0.5
1.0
1.5
2.0
0
0.5
1
0
0.5
1

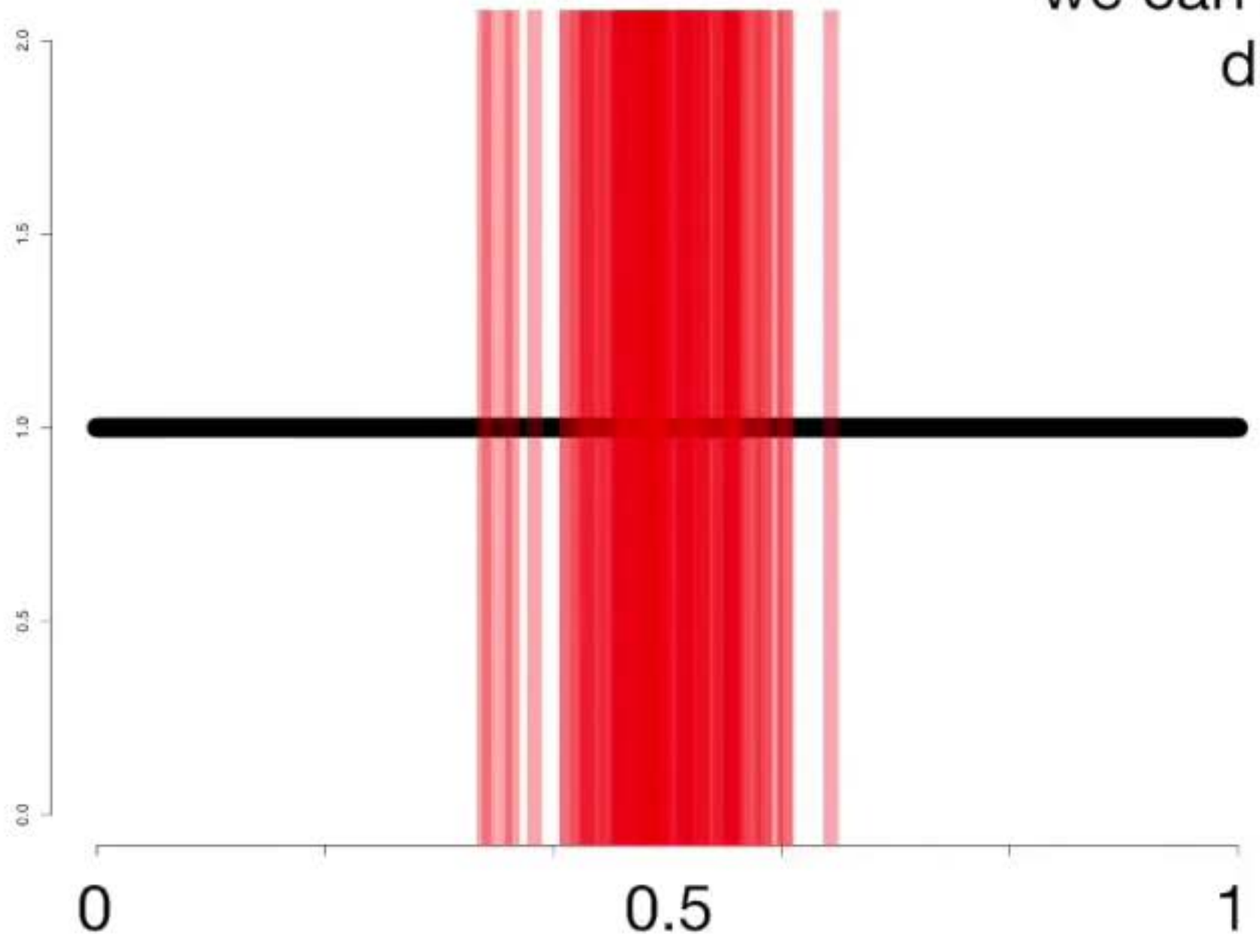

…100 means…
2.0
1.5
1.0
0.5
0.0
0
0.5
1
0
0.5
1

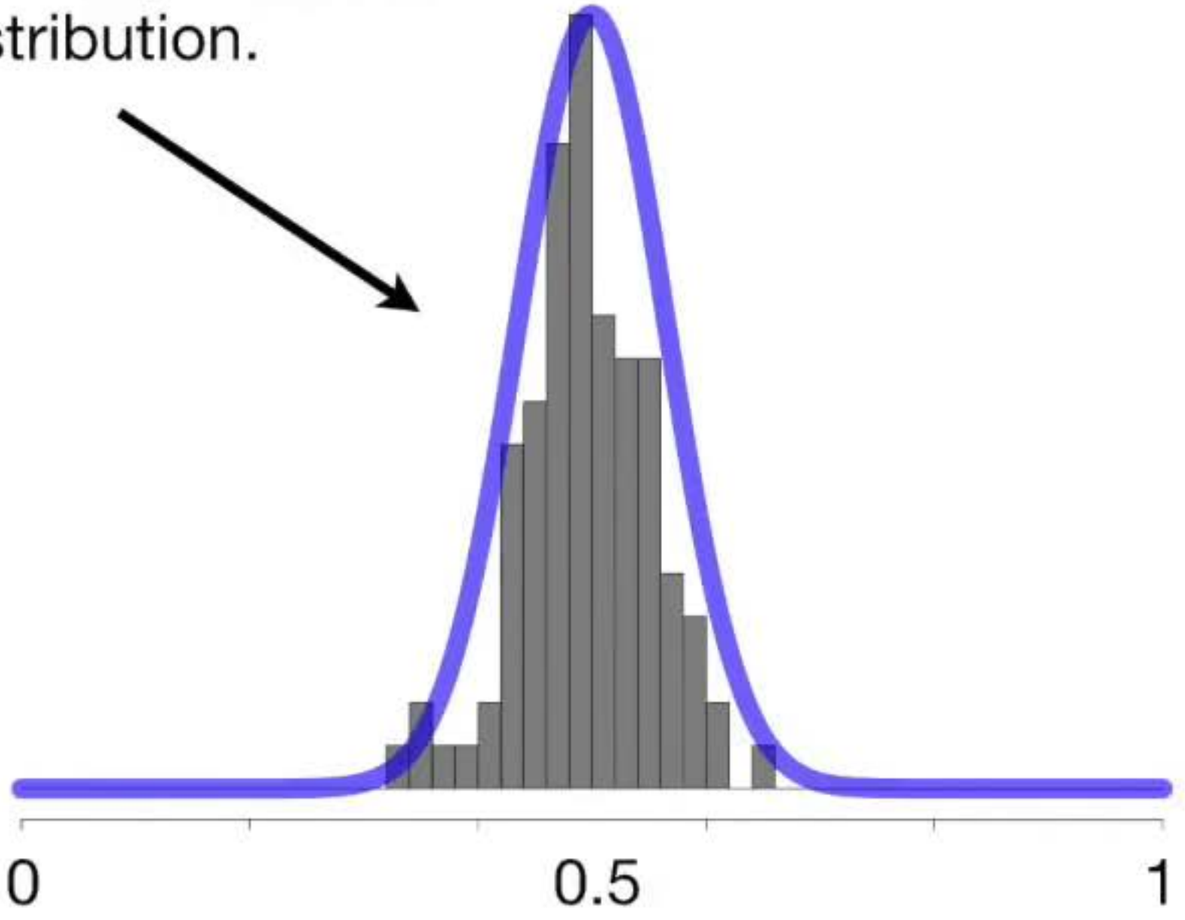
After adding 100 means to the histogram, it's pretty easy to see that these **means are normally distributed**.





However, to make super easy to see that the **means are normally distributed**, we can overlay a normal distribution.





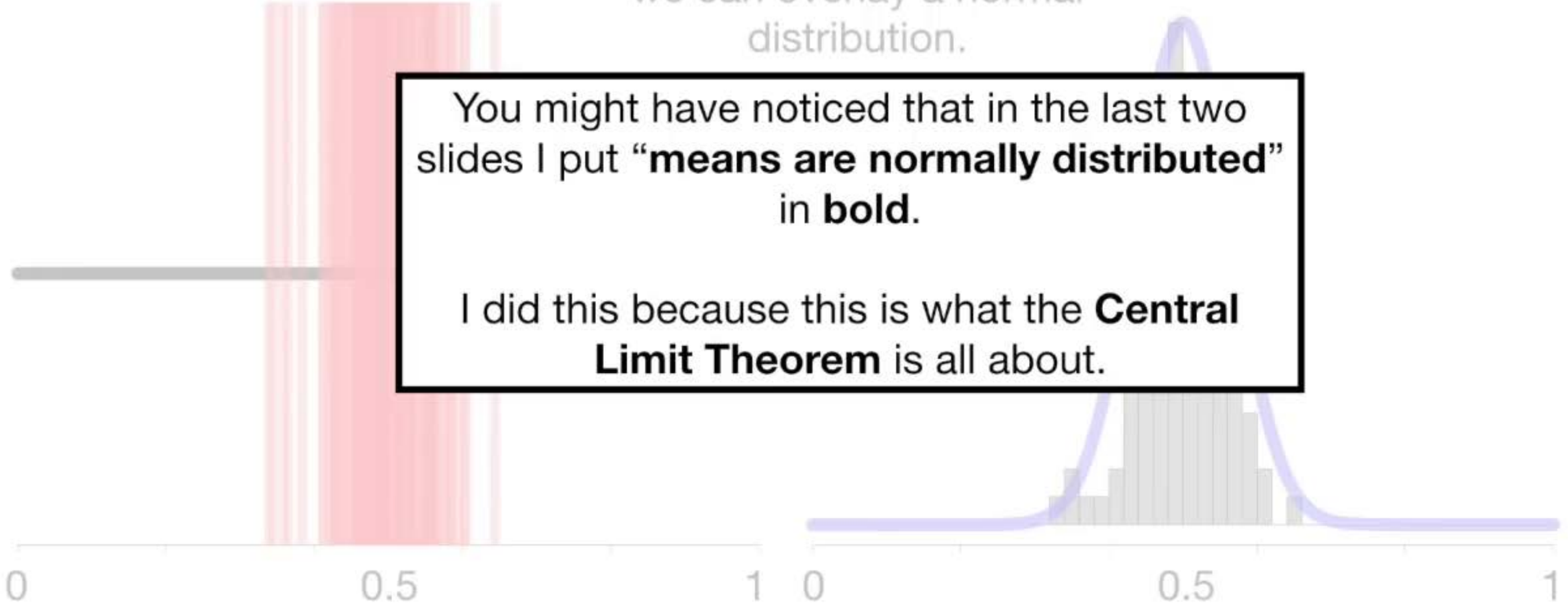
However, to make super easy to see that the

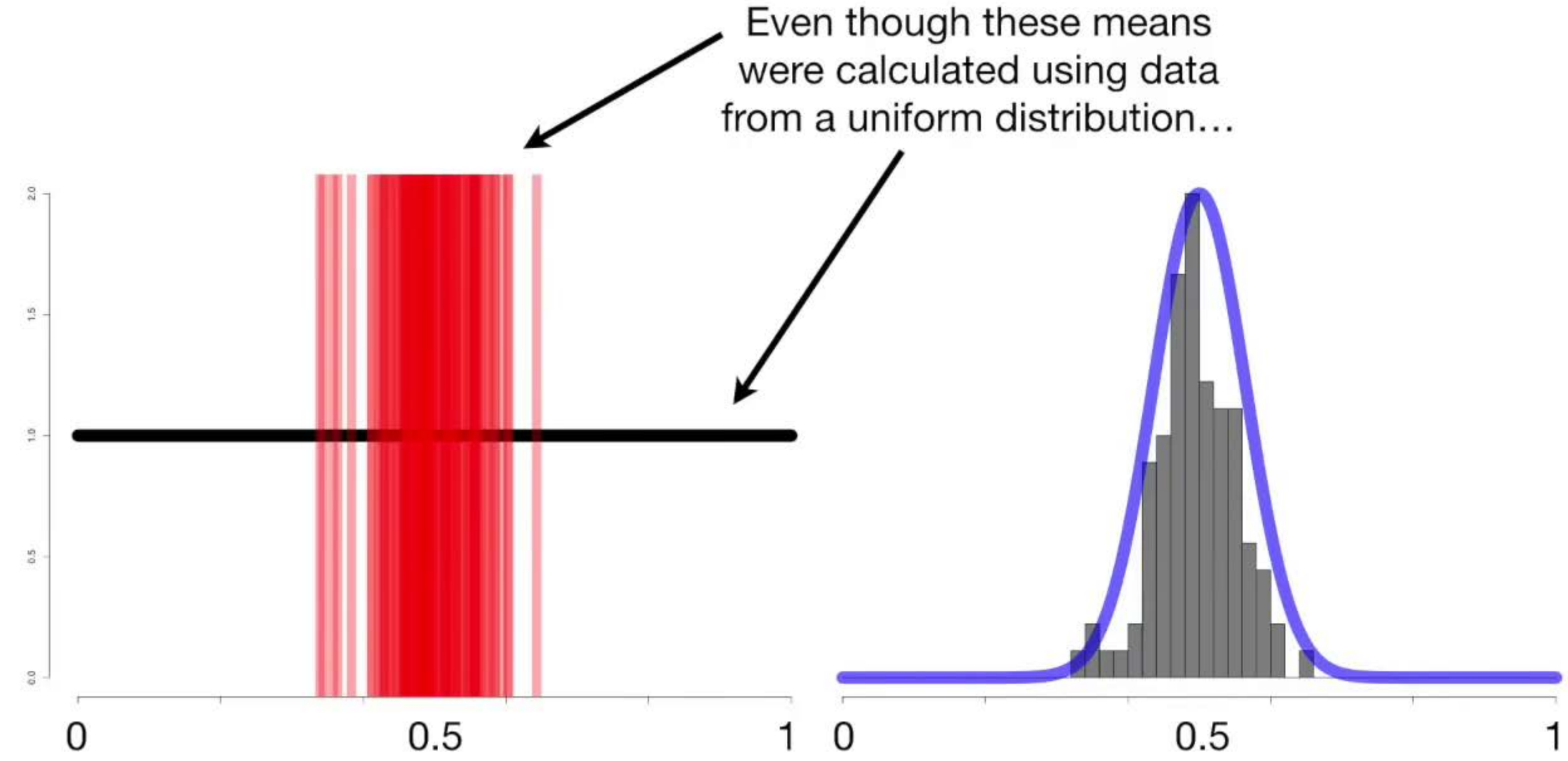
means are normally distributed,

we can overlay a normal distribution.

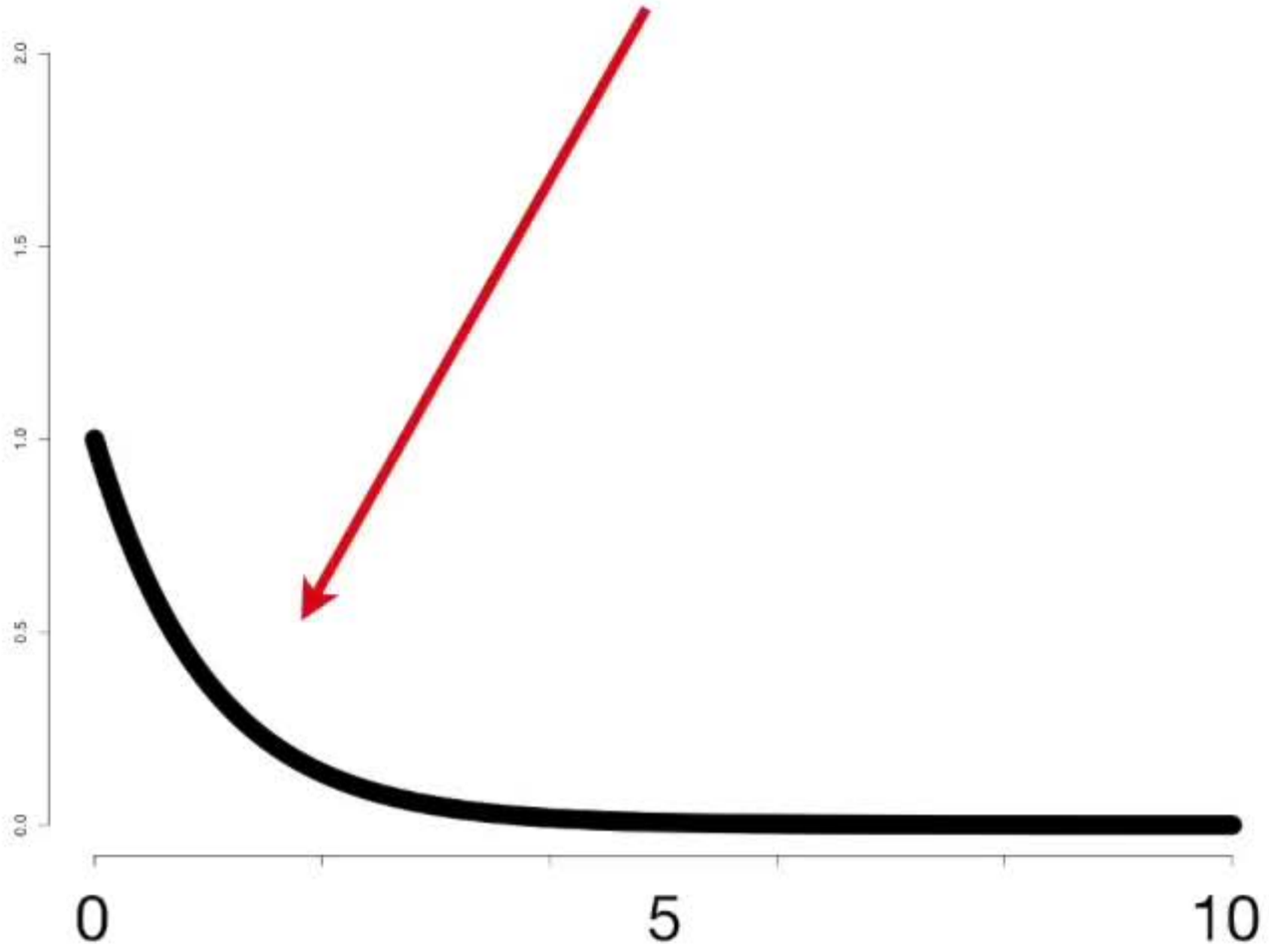
You might have noticed that in the last two slides I put "**means are normally distributed**" in **bold**.

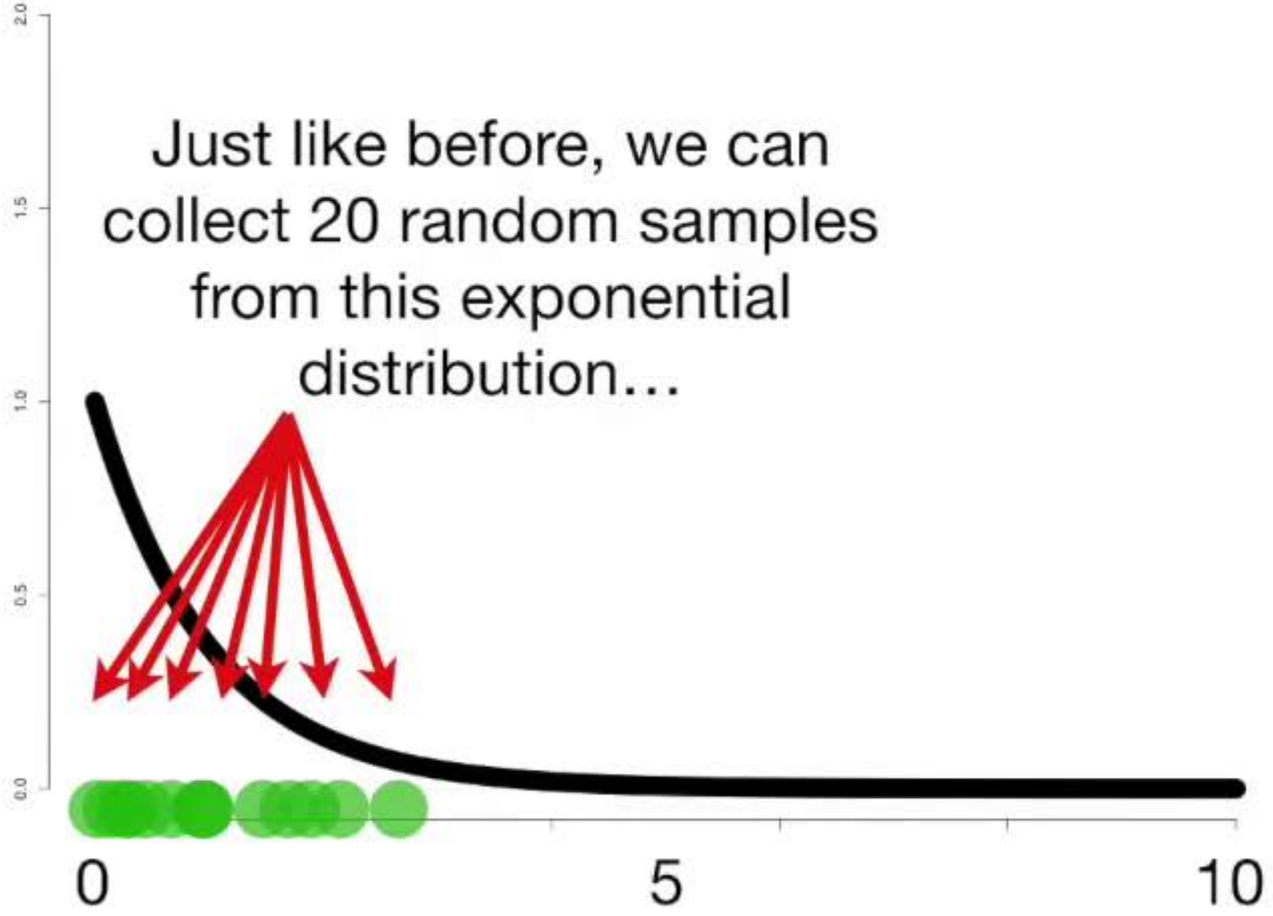
I did this because this is what the **Central Limit Theorem** is all about.

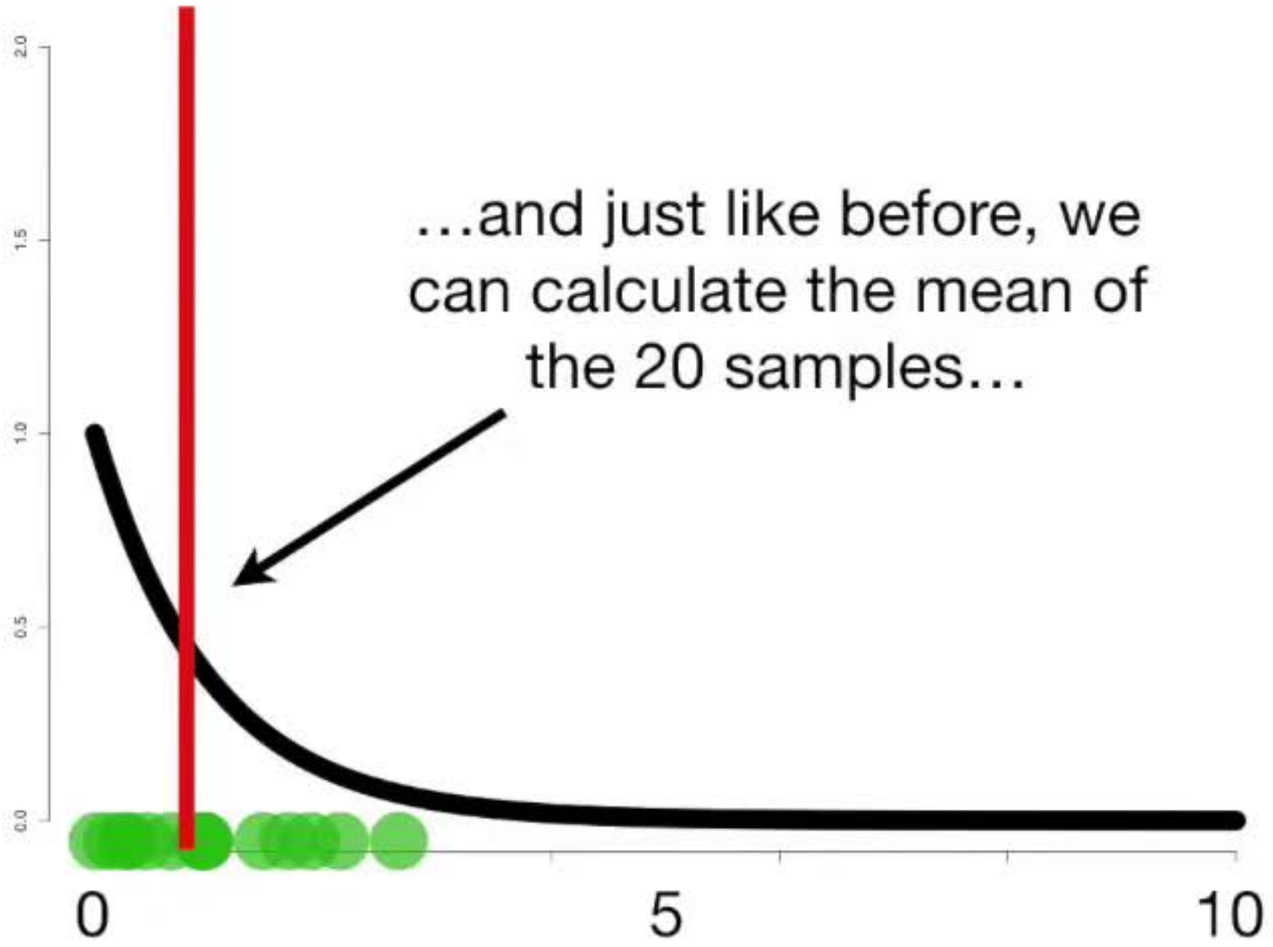



Even though these means
were calculated using data
from a uniform distribution…
0.0
0.5
1.0
1.5
2.0
0
0.5
1
0
0.5
1

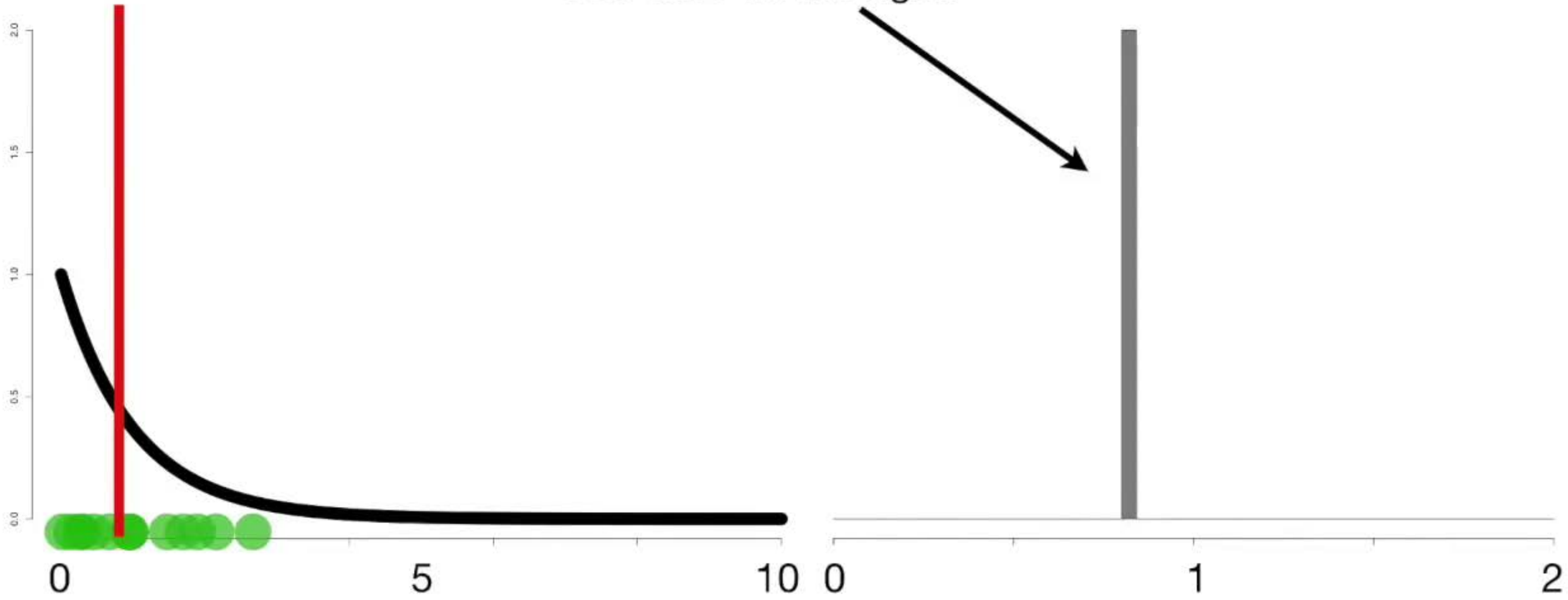
This time we'll start with an Exponential Distribution.



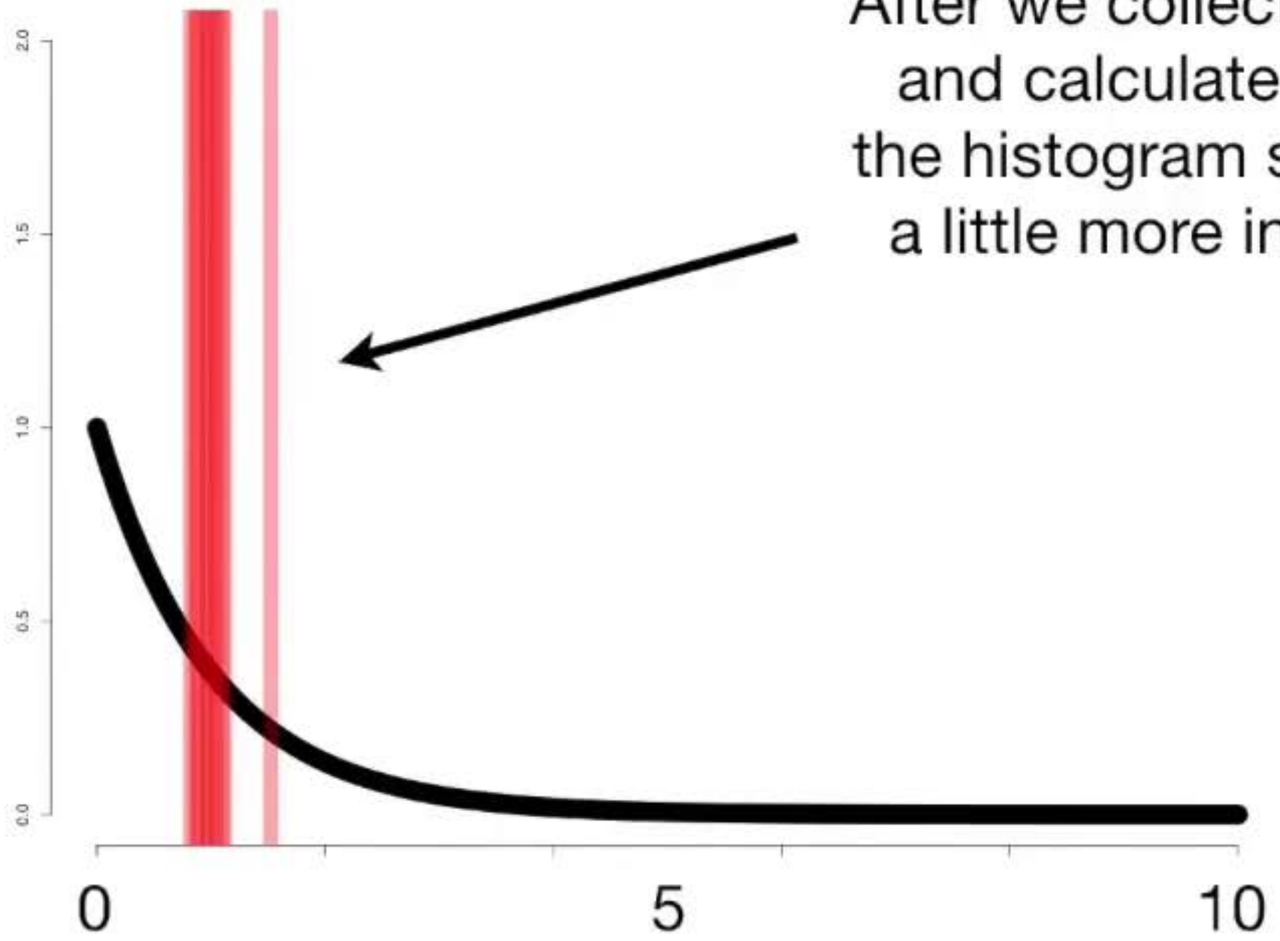

Just like before, we can collect 20 random samples from this exponential distribution…
0.0
0.5
1.0
1.5
2.0
0
5
10

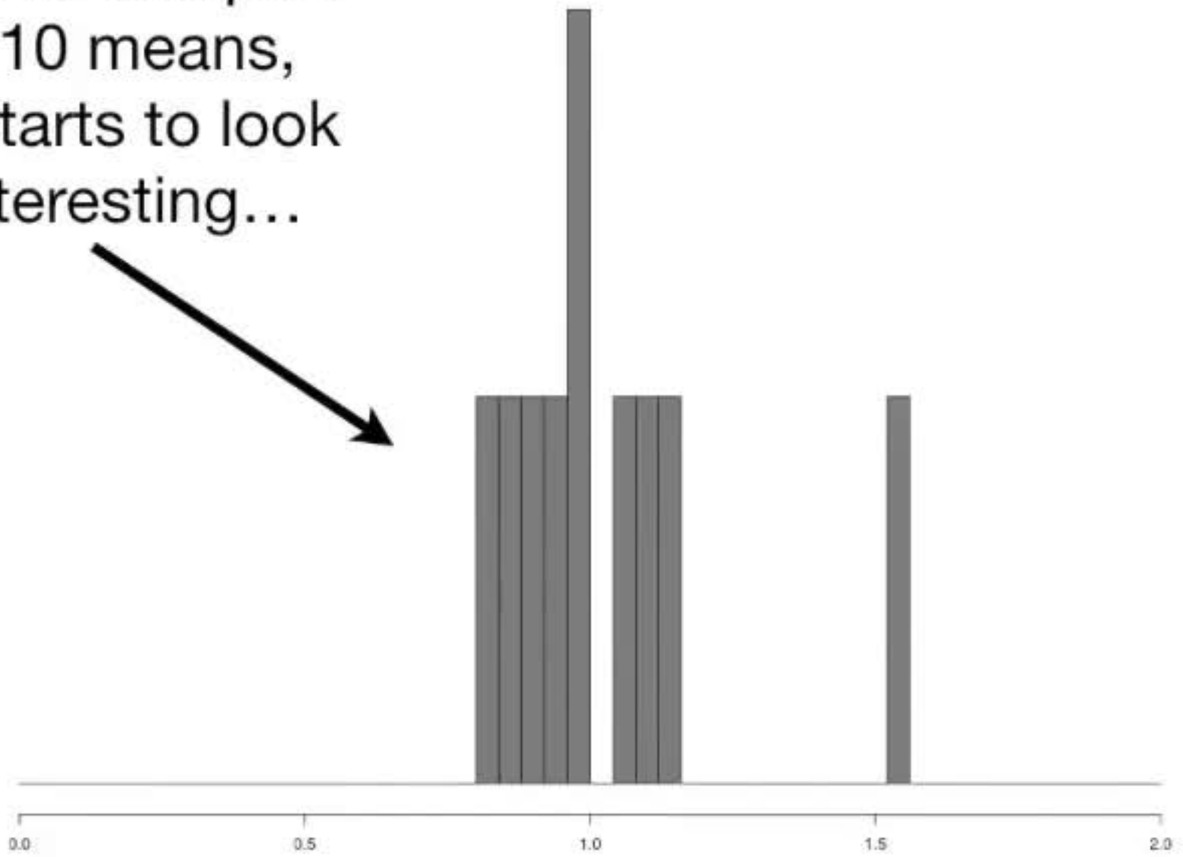

…and just like before, we
can calculate the mean of
the 20 samples…
0.0
0.5
1.0
1.5
2.0
0
5
10

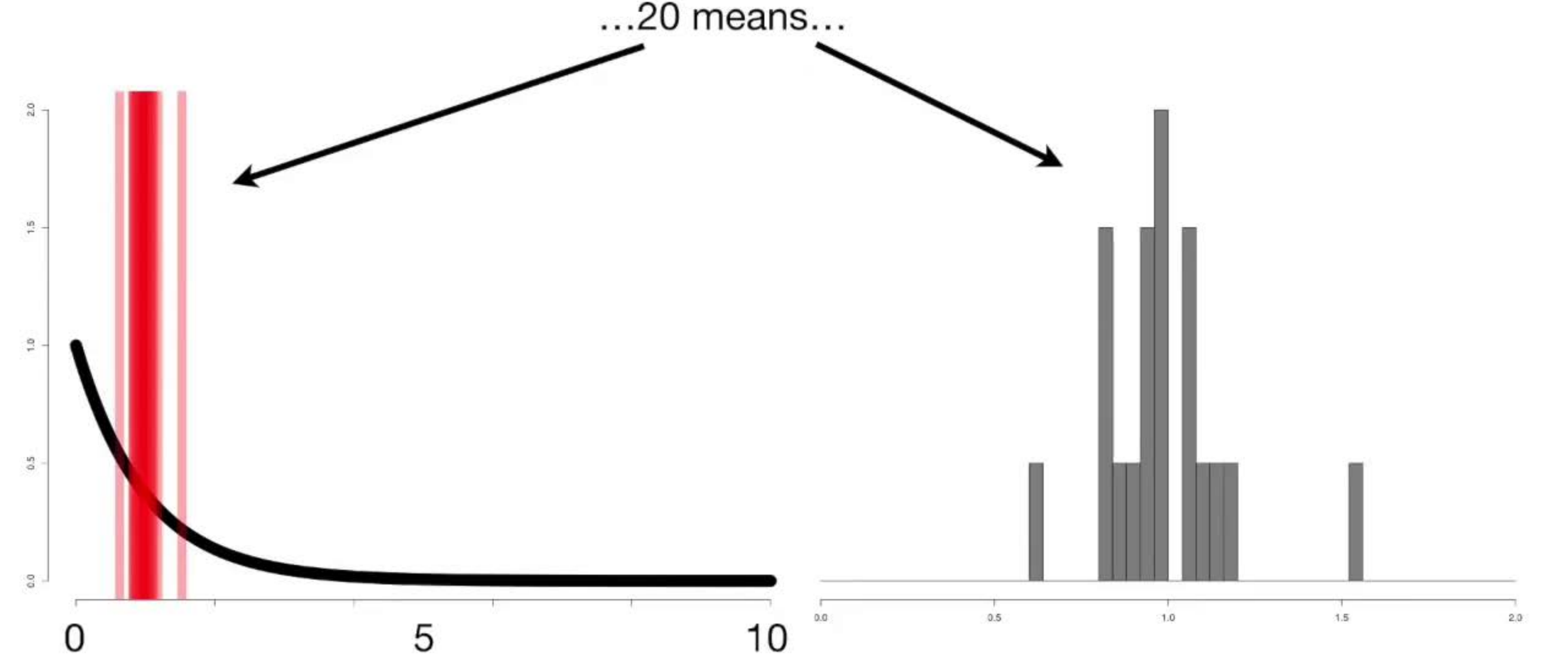
…and lastly, we can draw a
histogram of that mean
over here on the right.

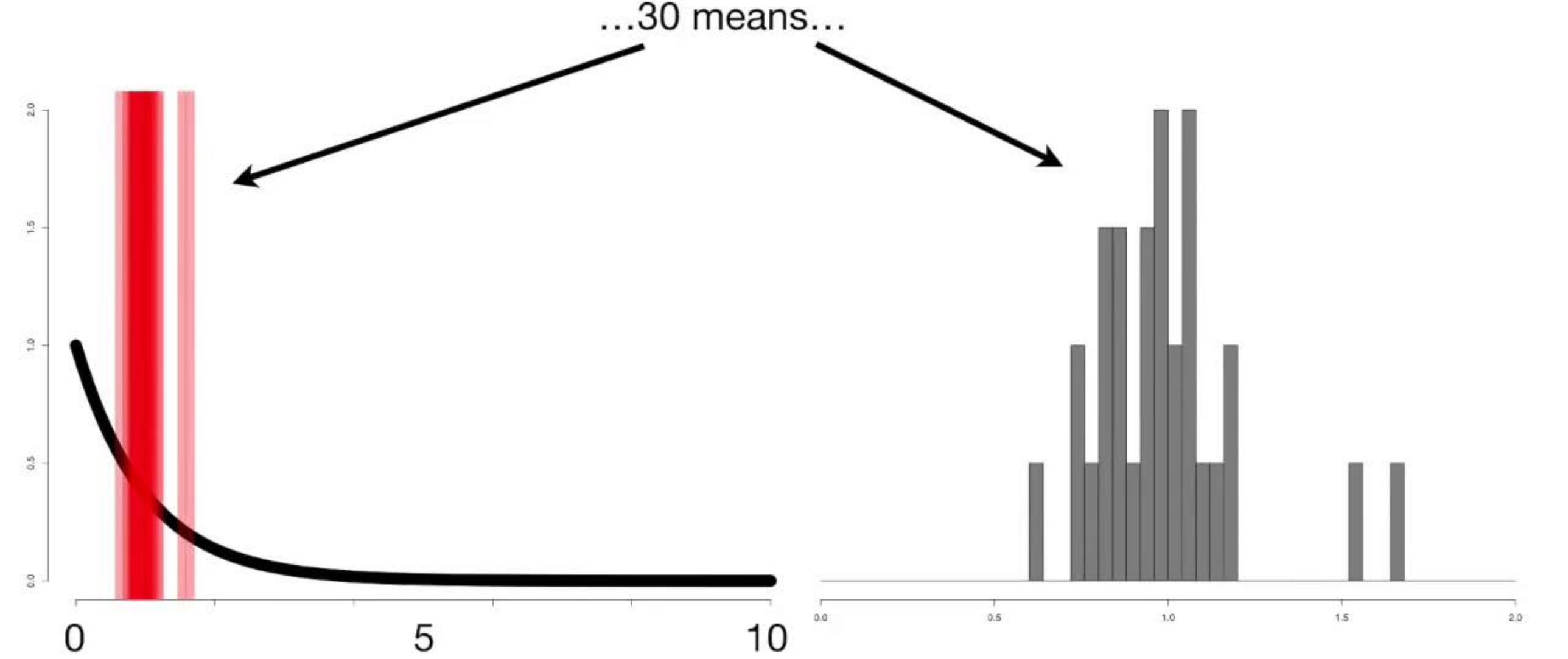


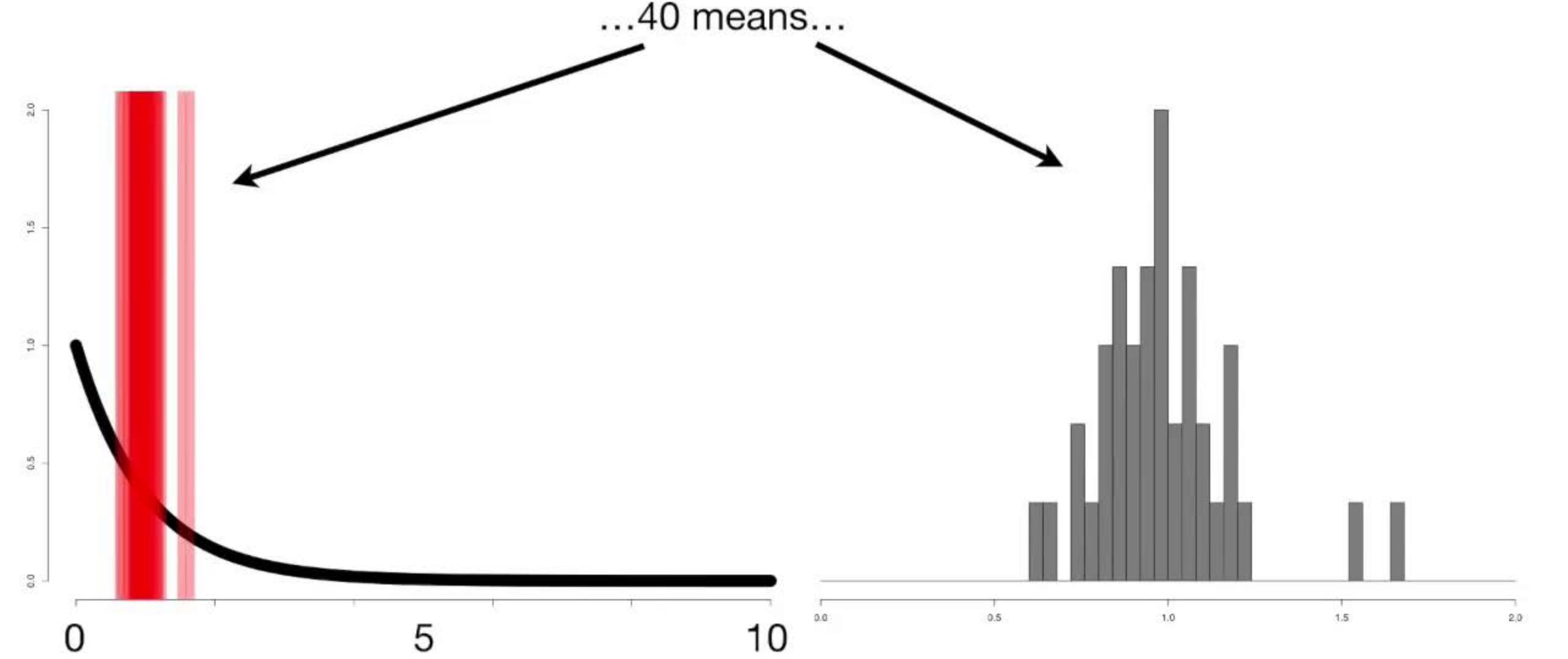
After we collect 10 samples
and calculate 10 means,
the histogram starts to look
a little more interesting…

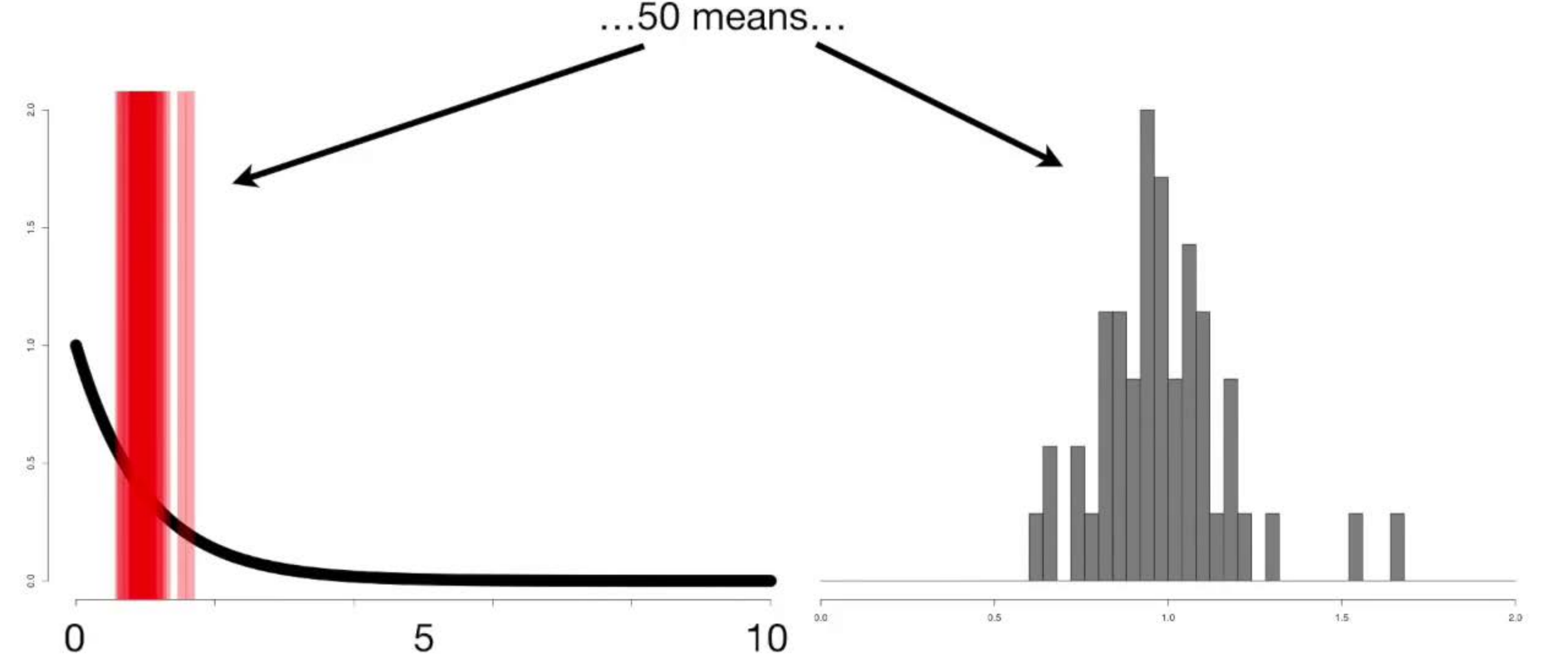


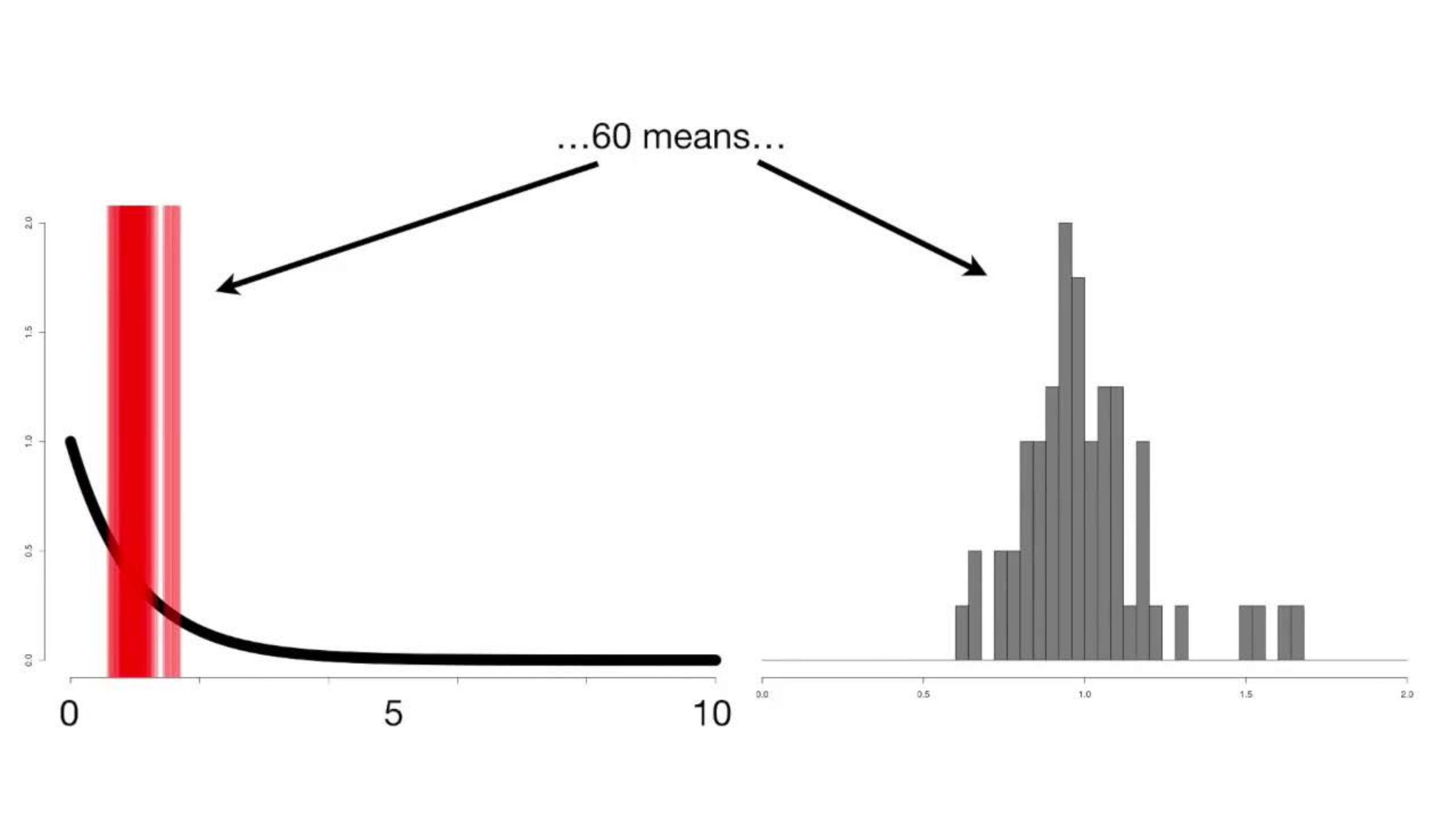


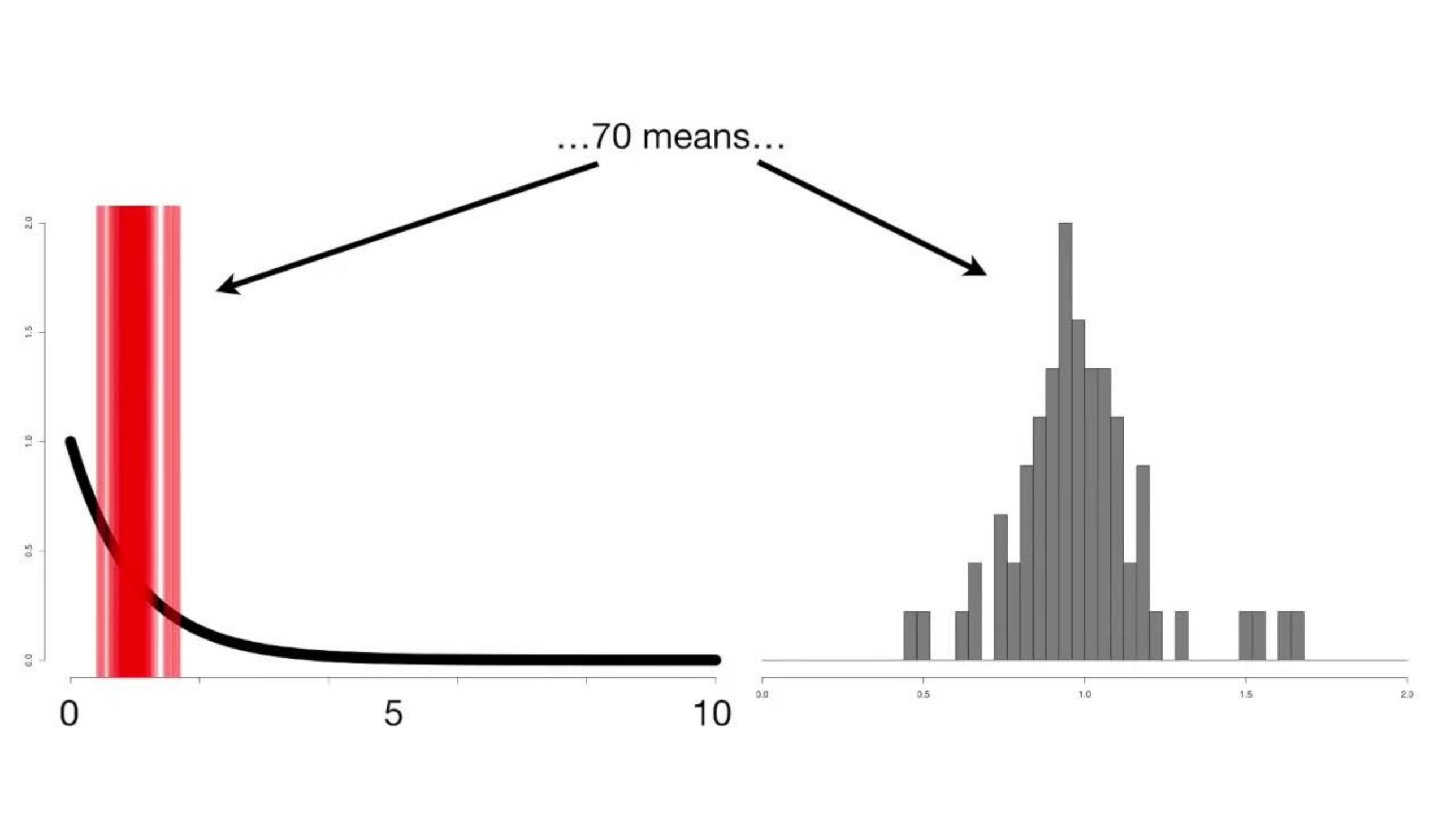

…20 means…
0.0
0.5
1.0
1.5
2.0
0
5
10
0.0
0.5
1.0
1.5
2.0

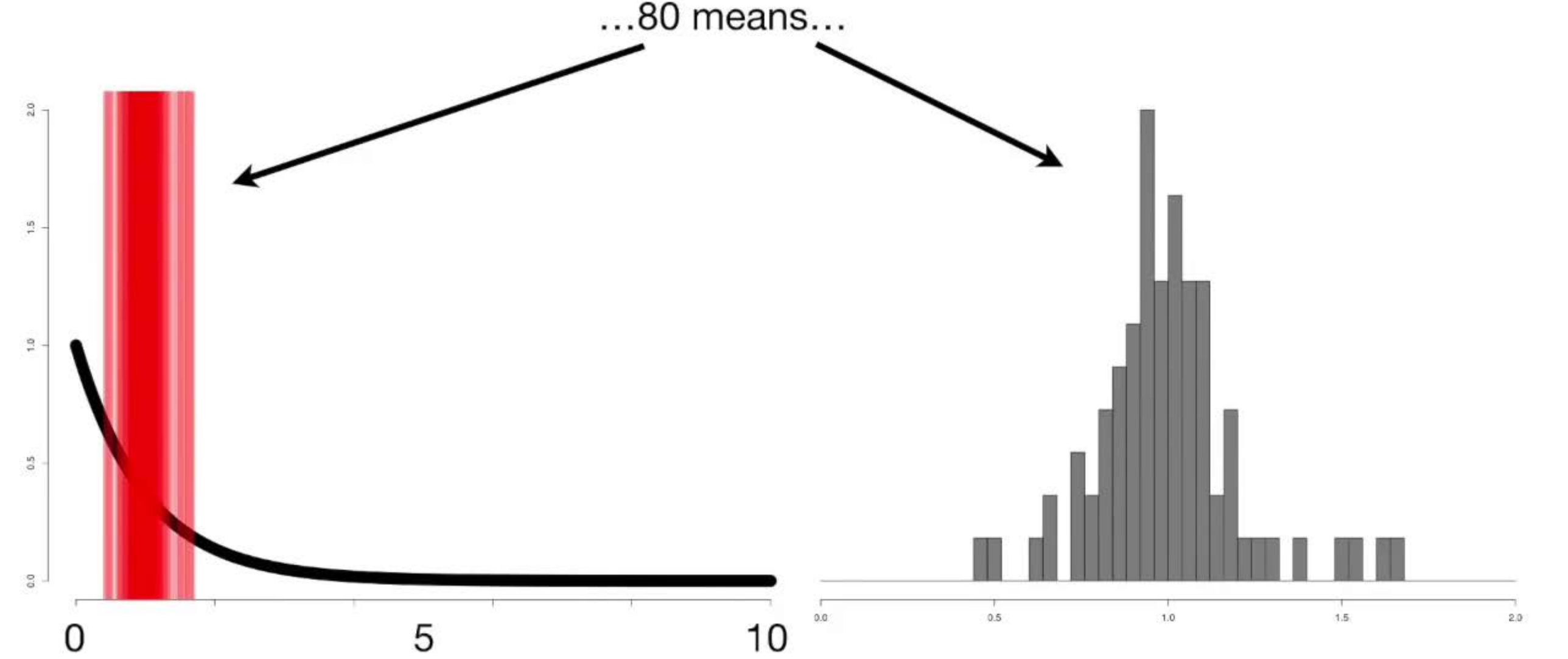

…30 means…
0.0
0.5
1.0
1.5
2.0
0
5
10
0.0
0.5
1.0
1.5
2.0

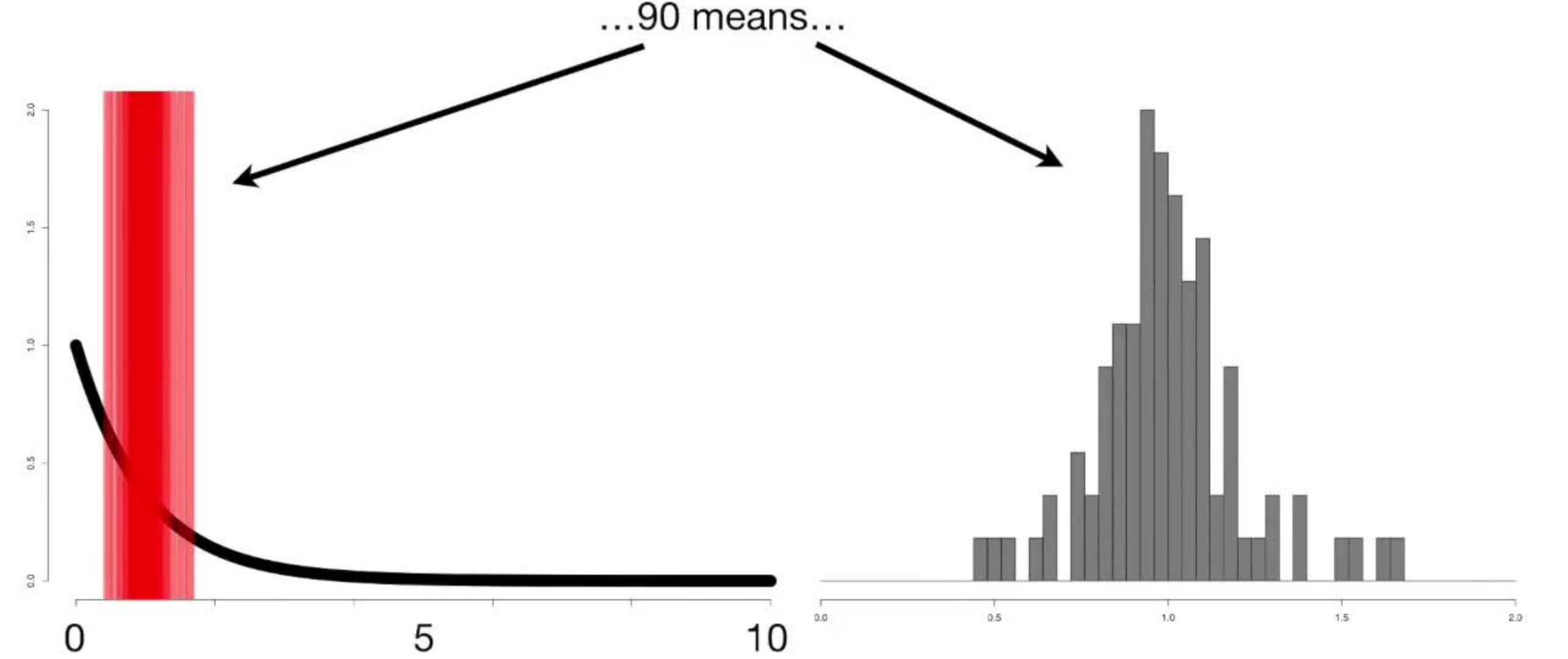

…40 means…
0.0
0.5
1.0
1.5
2.0
0
5
10
0.0
0.5
1.0
1.5
2.0

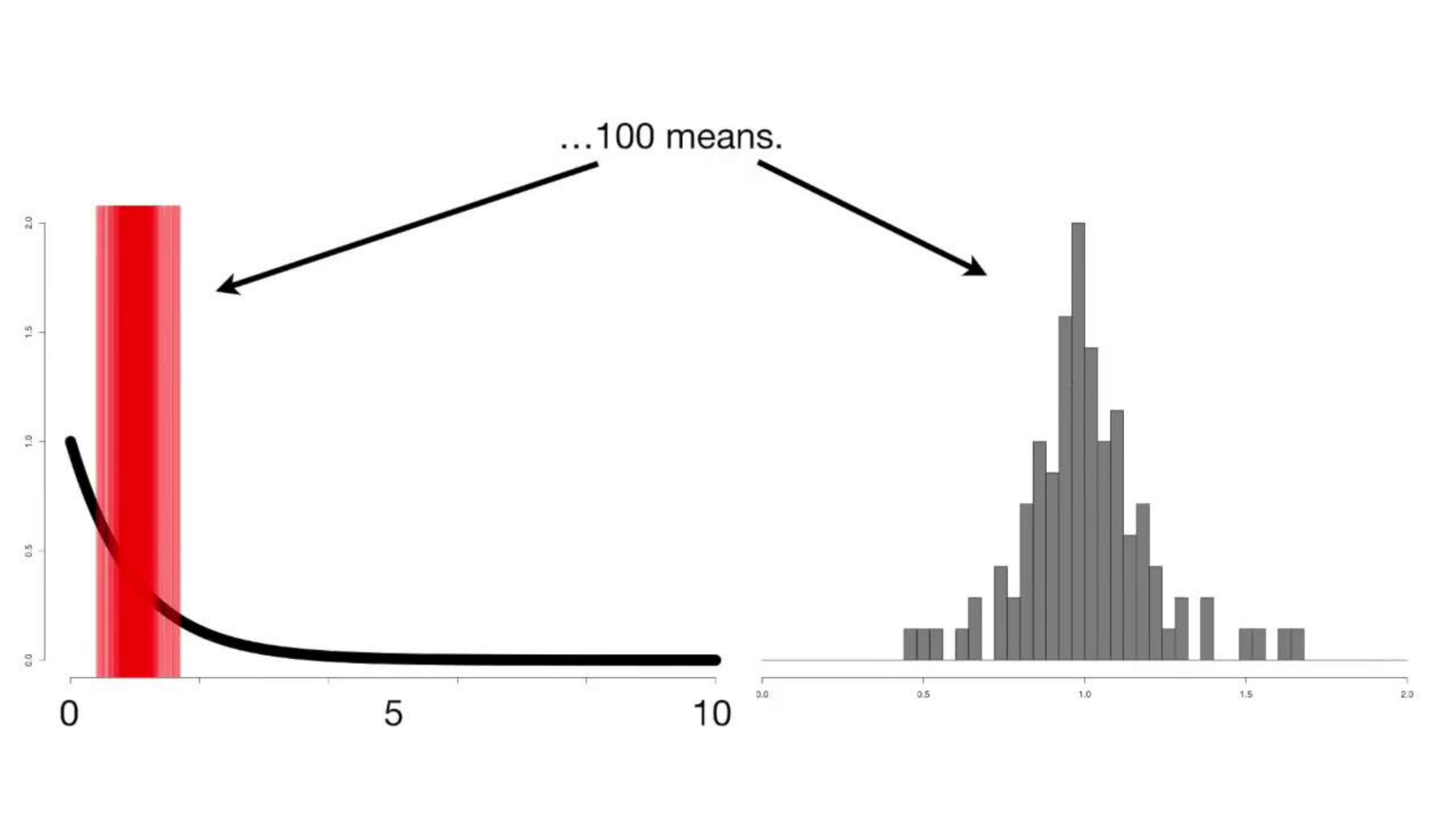

…50 means…
0.0
0.5
1.0
1.5
2.0
0
5
10
0.0
0.5
1.0
1.5
2.0


…60 means…
0.0
0.5
1.0
1.5
2.0
0
5
10
0.0
0.5
1.0
1.5
2.0

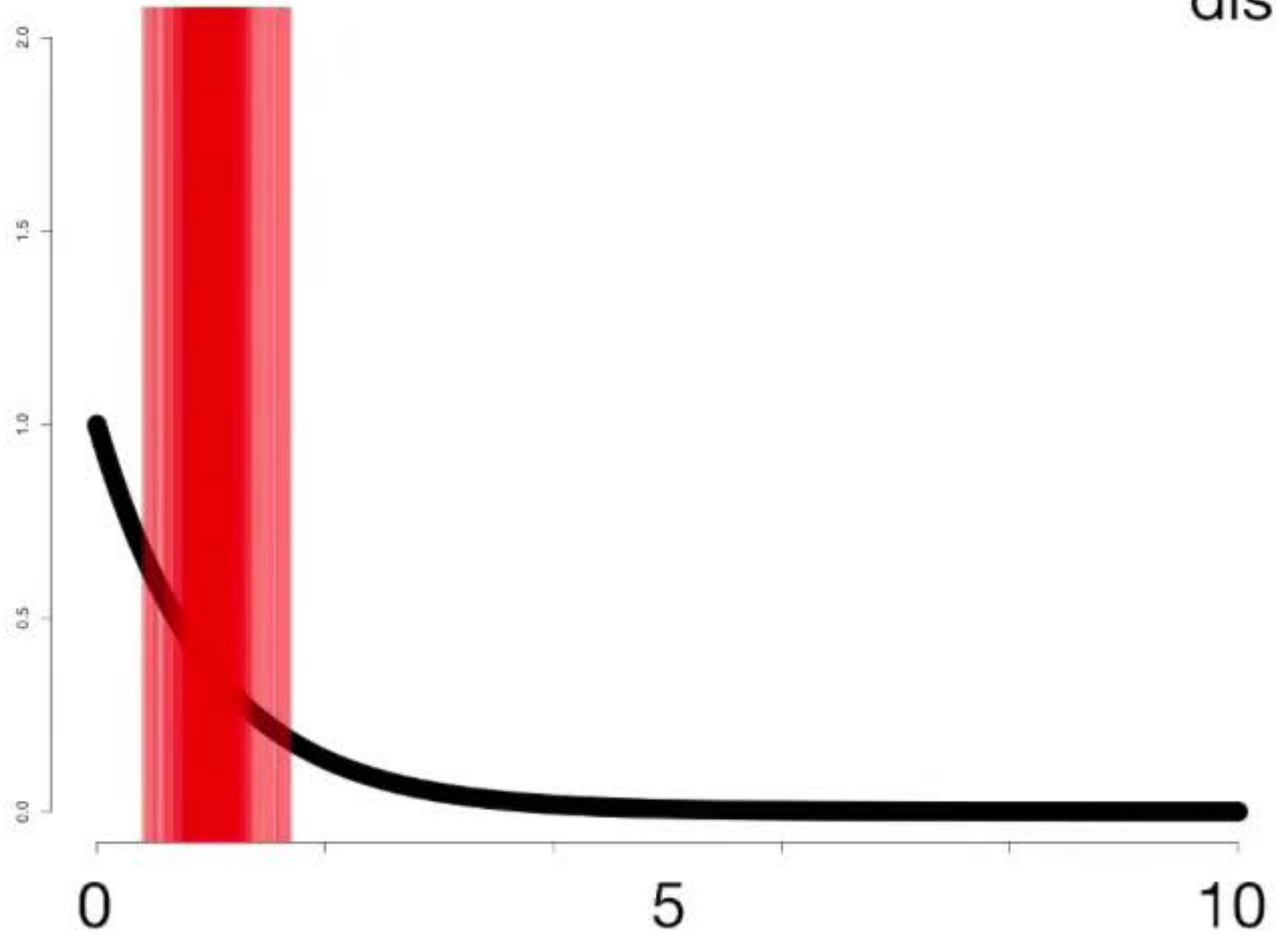

…70 means…
0.0
0.5
1.0
1.5
2.0
0
5
10
0.0
0.5
1.0
1.5
2.0

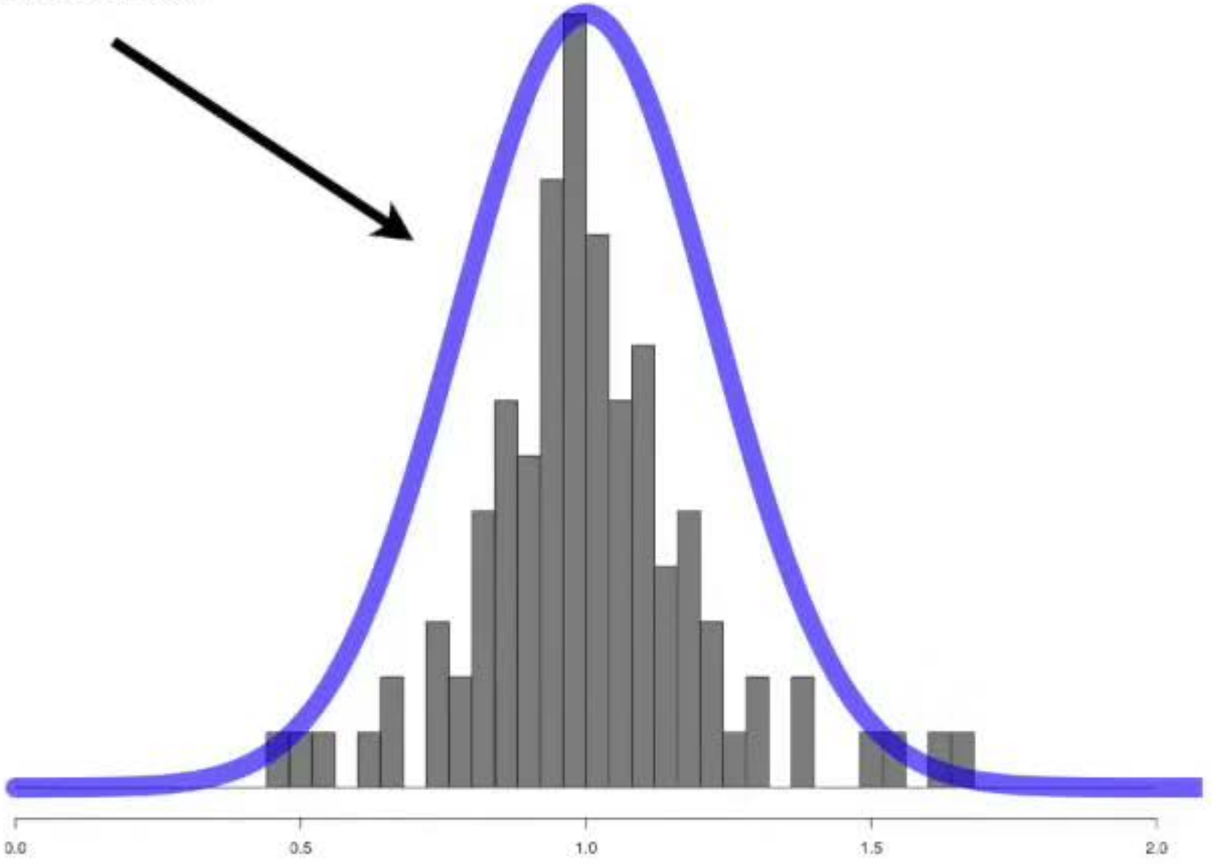

…80 means…
0.0
0.5
1.0
1.5
2.0
0
5
10
0.0
0.5
1.0
1.5
2.0


…90 means…
0.0
0.5
1.0
1.5
2.0
0
5
10
0.0
0.5
1.0
1.5
2.0

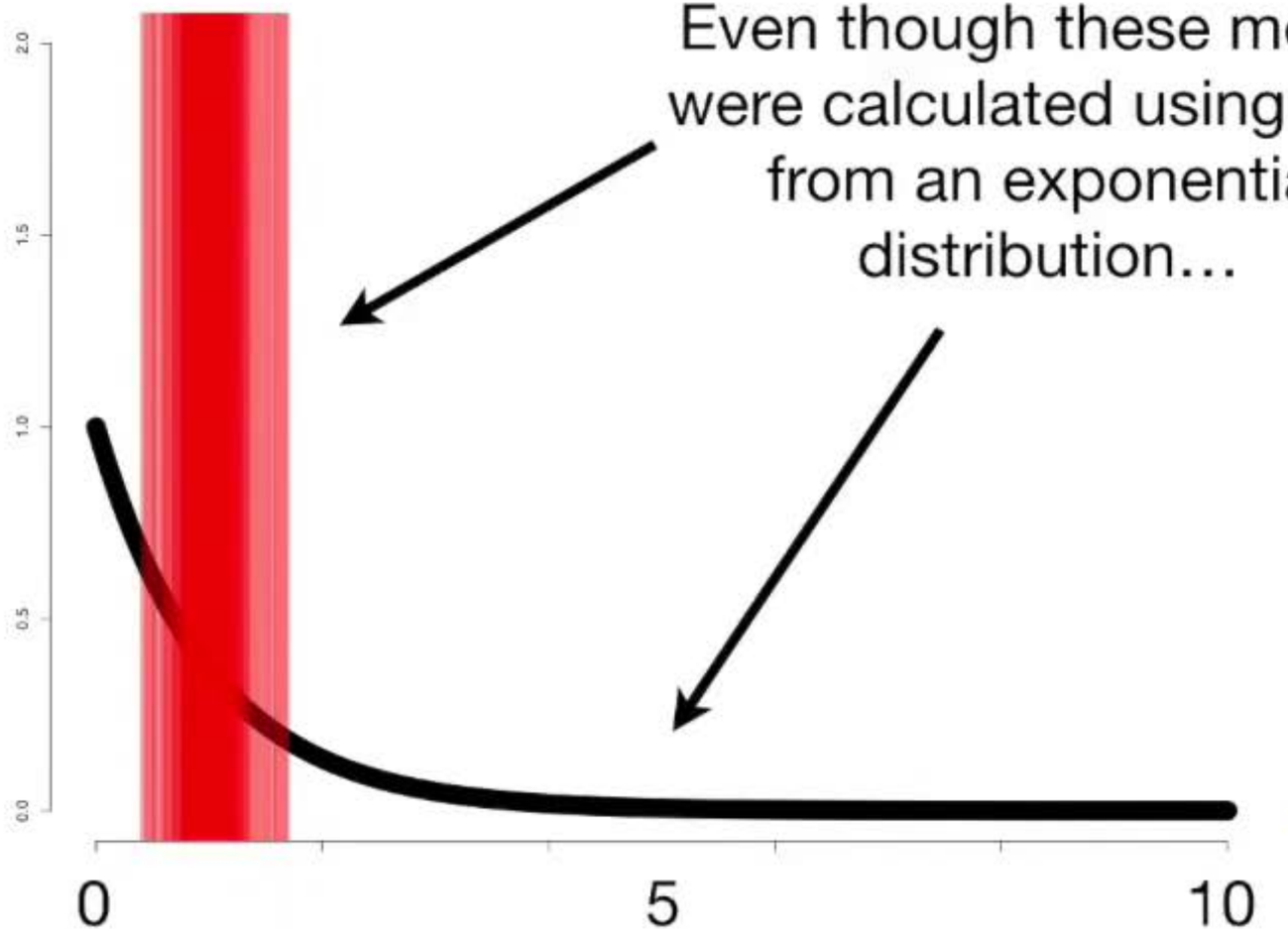

…100 means.
0.0
0.5
1.0
1.5
2.0
0
5
10
0.0
0.5
1.0
1.5
2.0

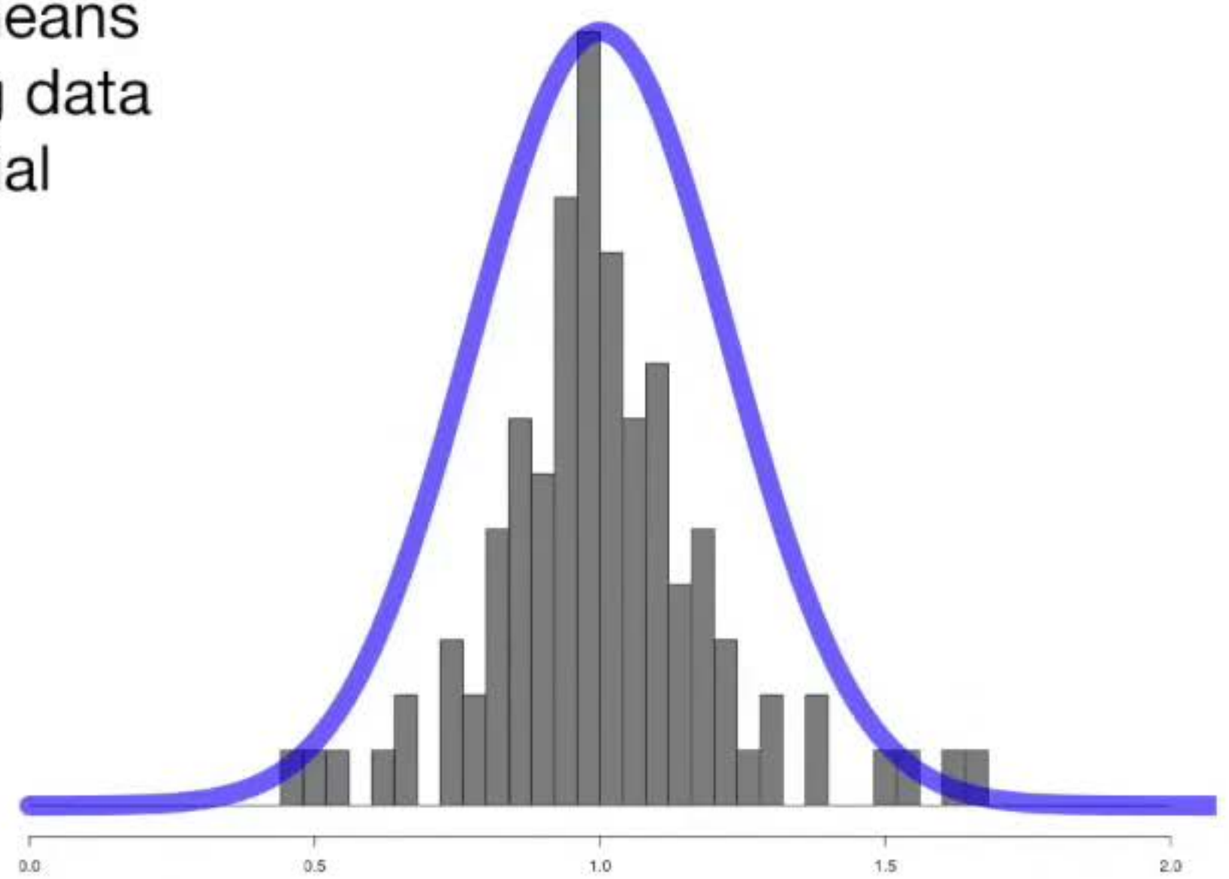
After adding 100 means to the histogram, we can see that they are normally distributed.



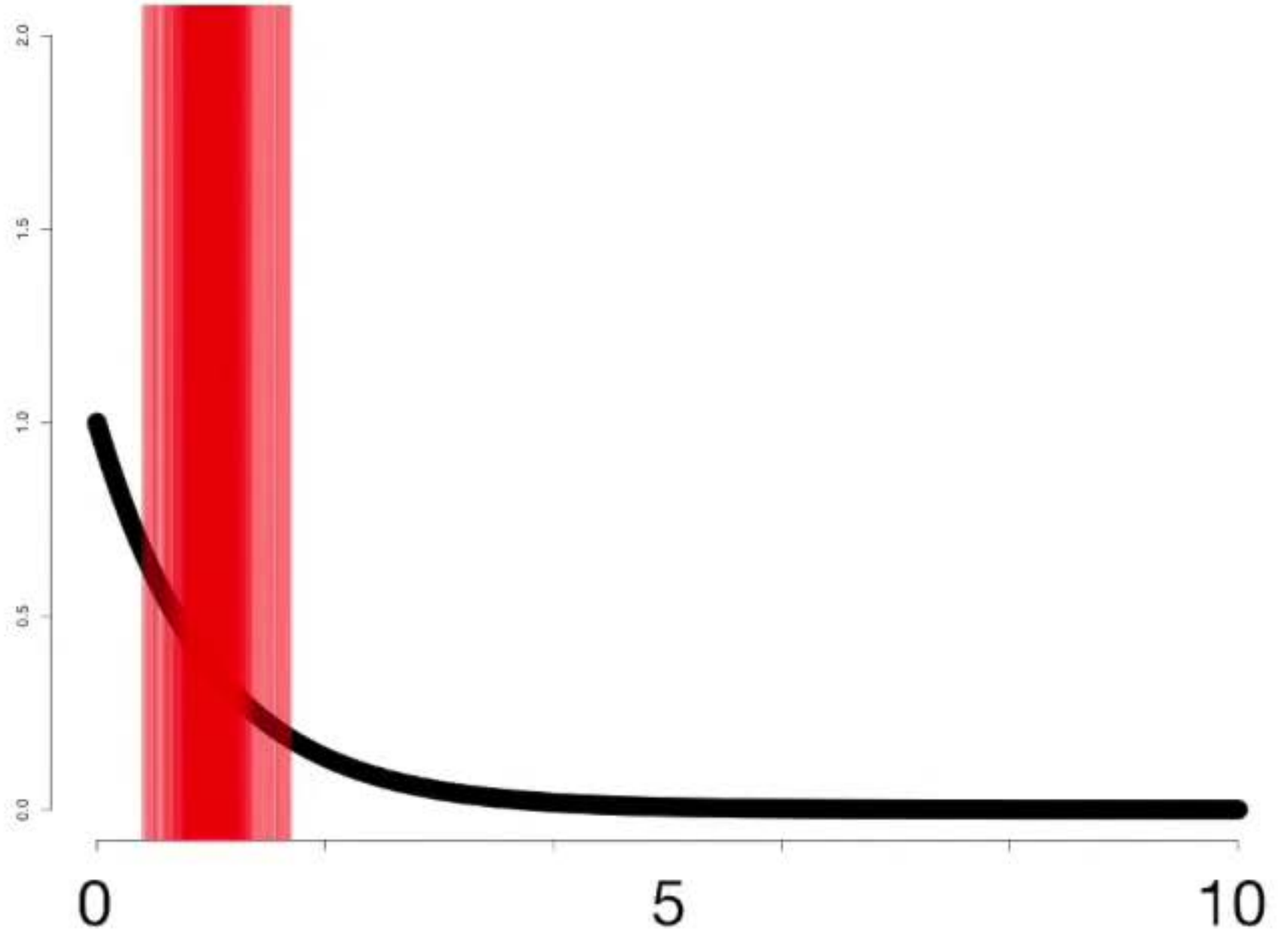


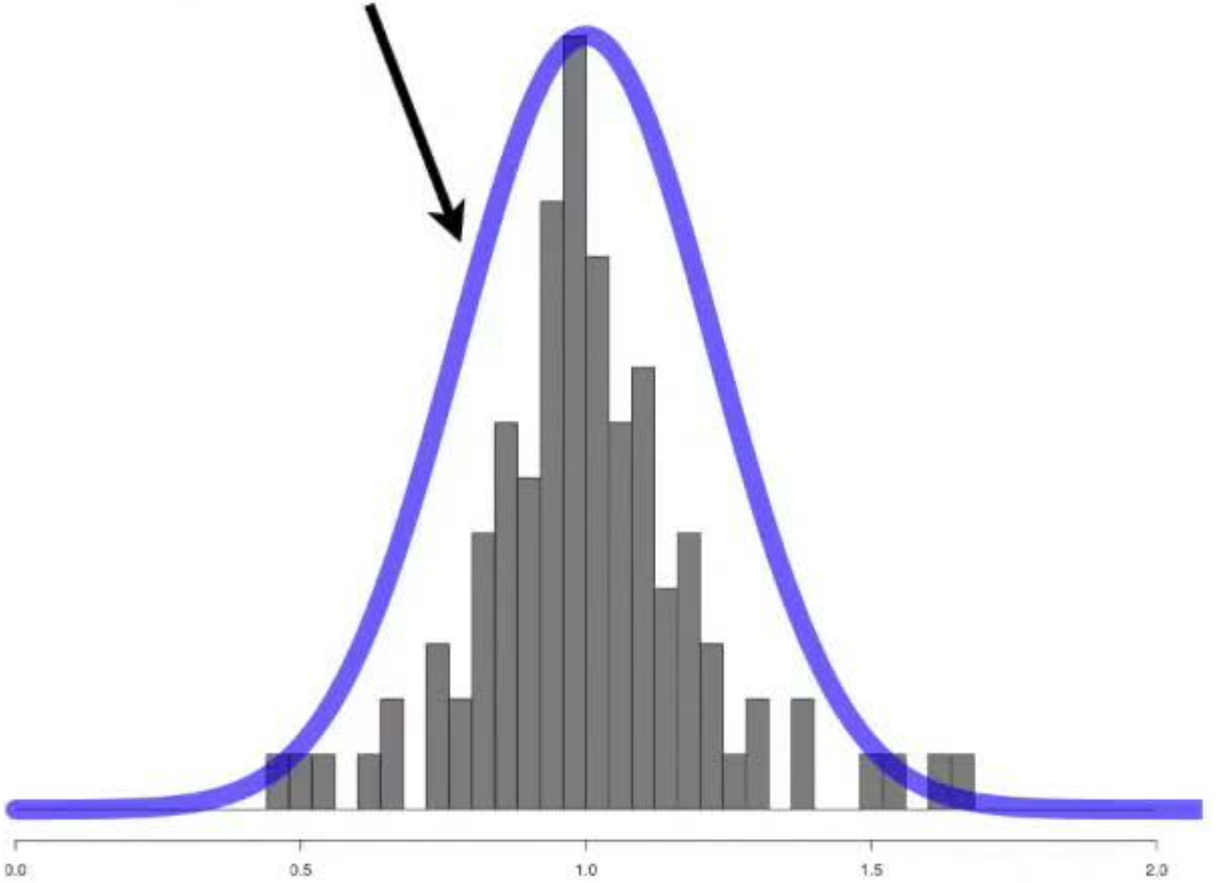
Even though these means were calculated using data from an exponential distribution…

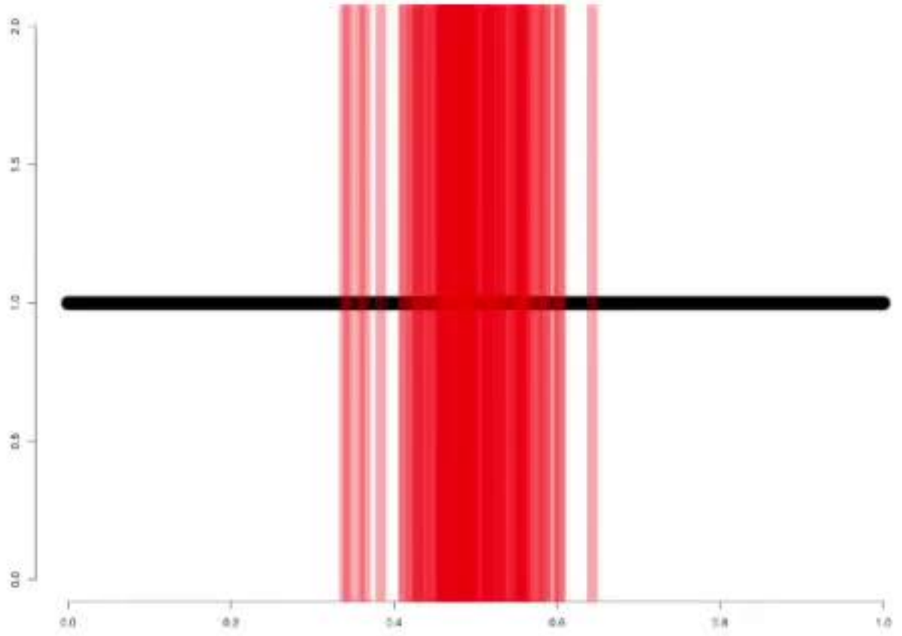




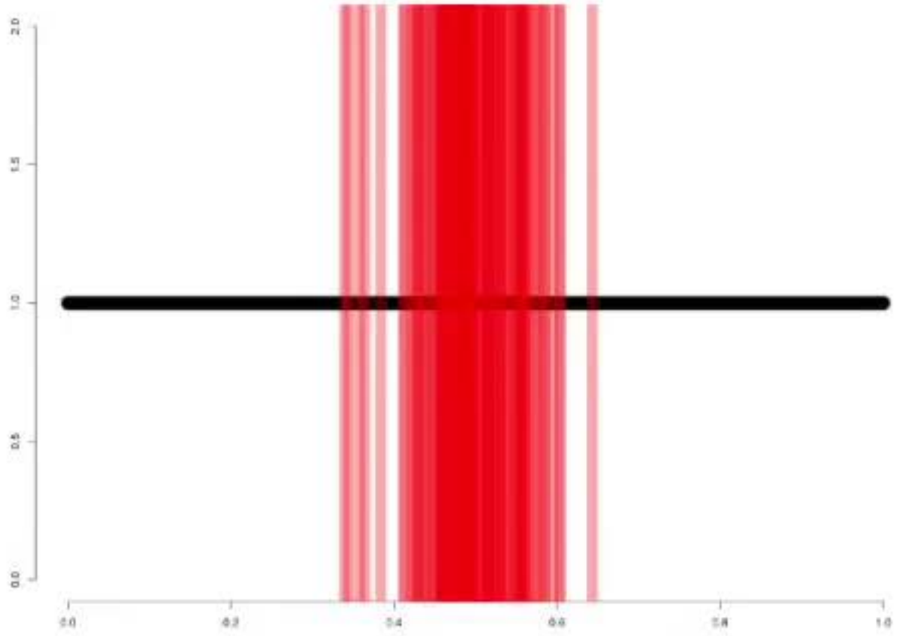
…the means themselves are not exponentially distributed. Instead, the **means are normally distributed**.



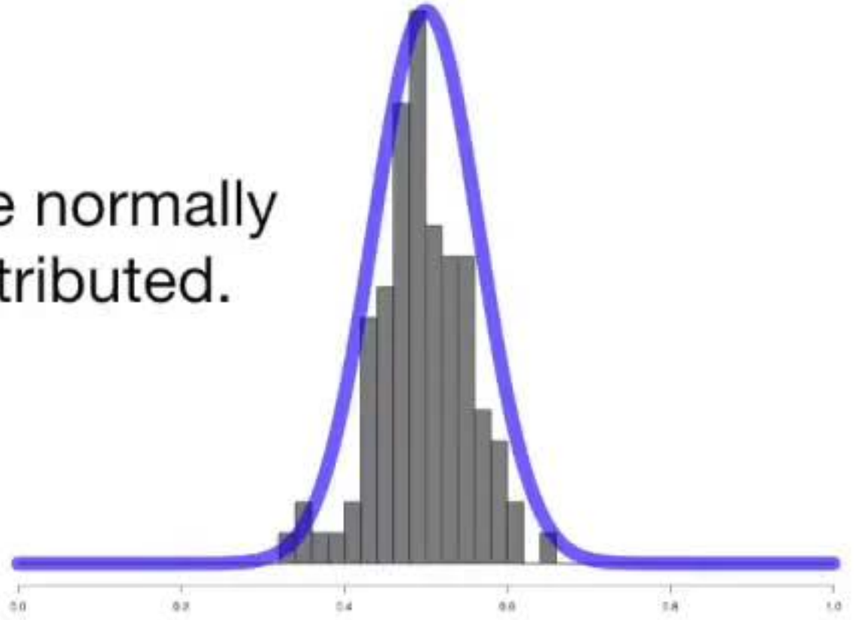


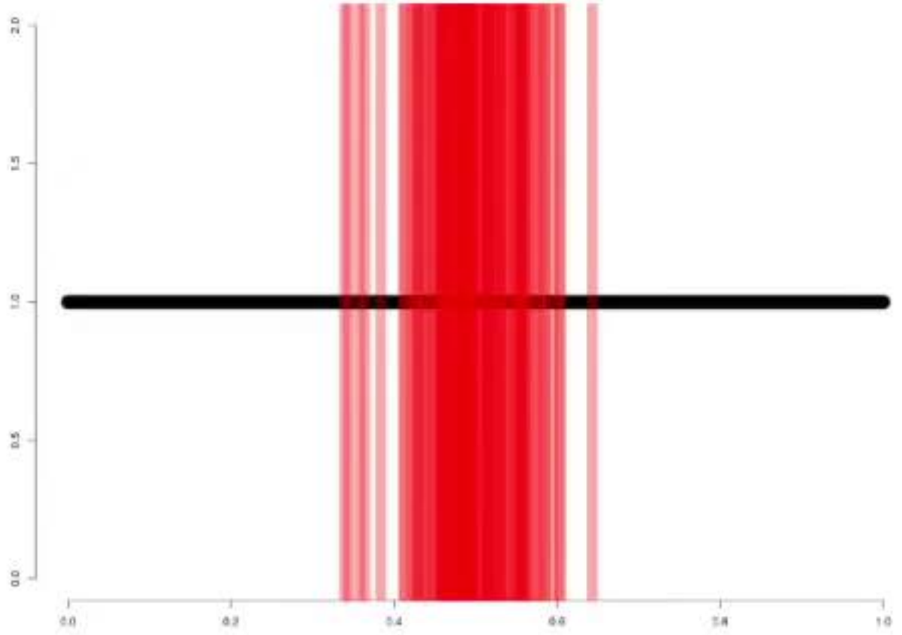


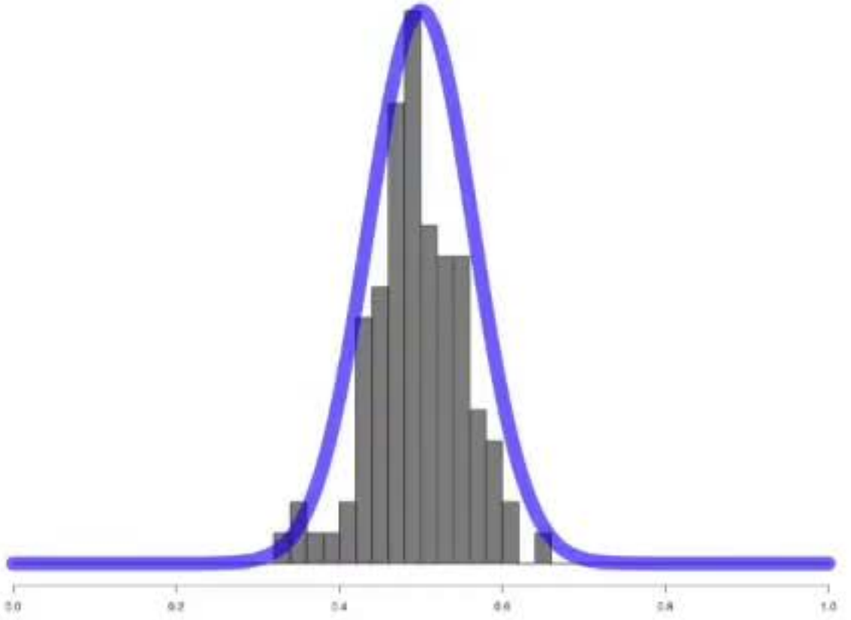
So far we have seen that means calculated from samples taken from a uniform distribution…

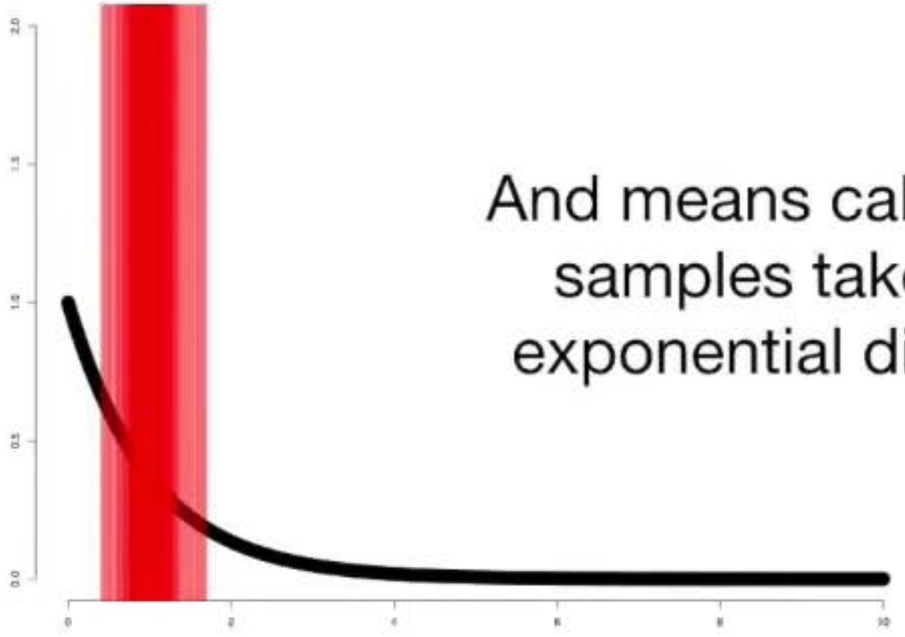


…are normally distributed.

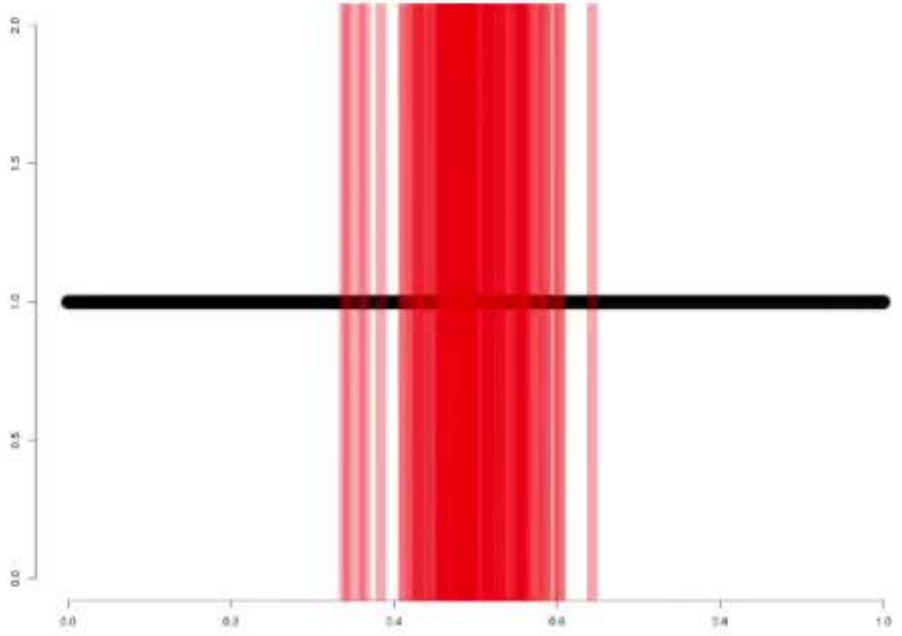


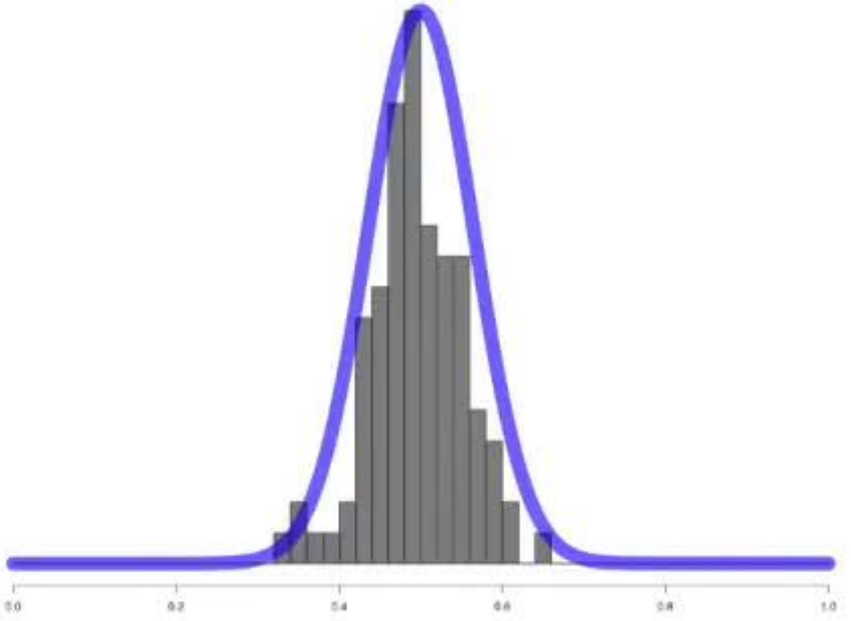


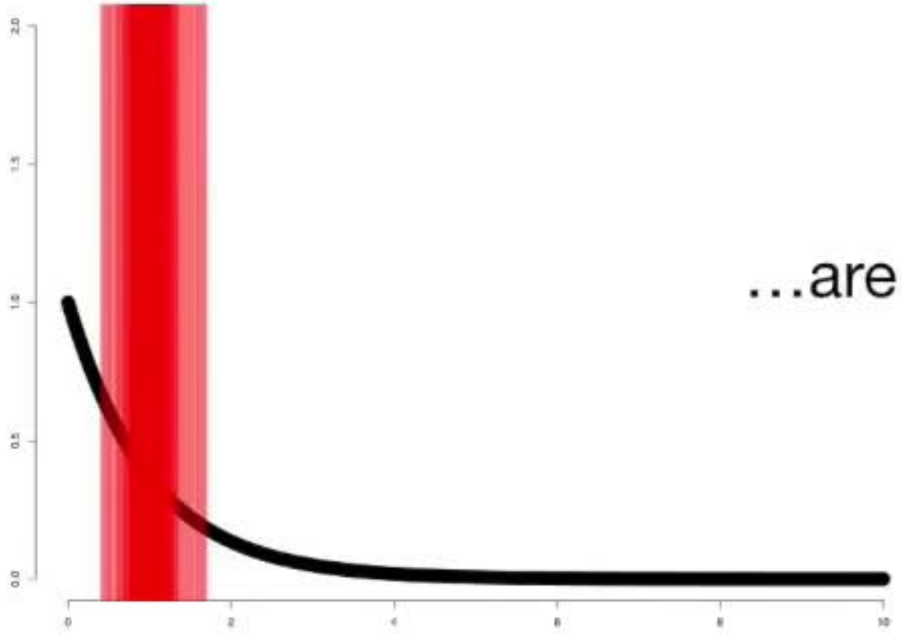




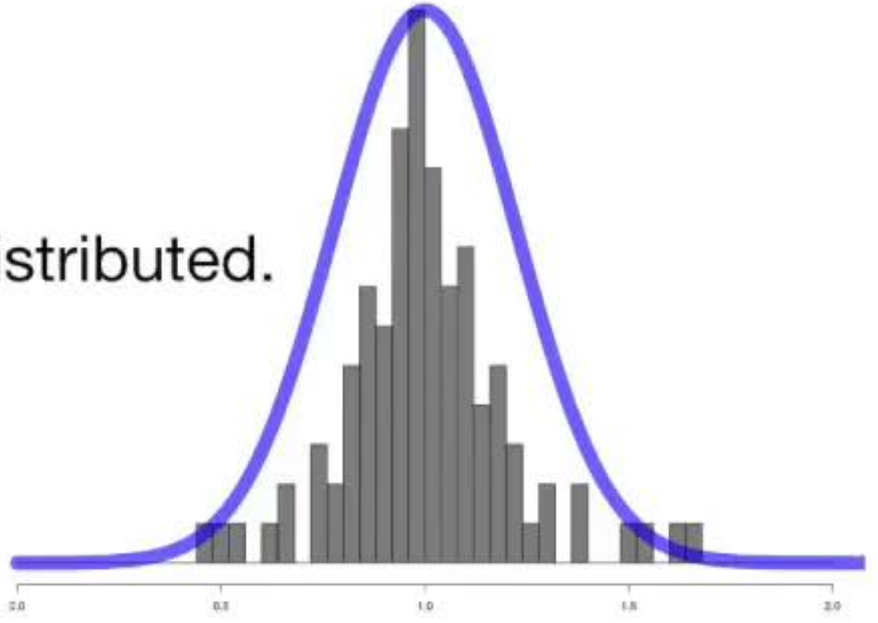
And means calculated from samples taken from an exponential distribution…



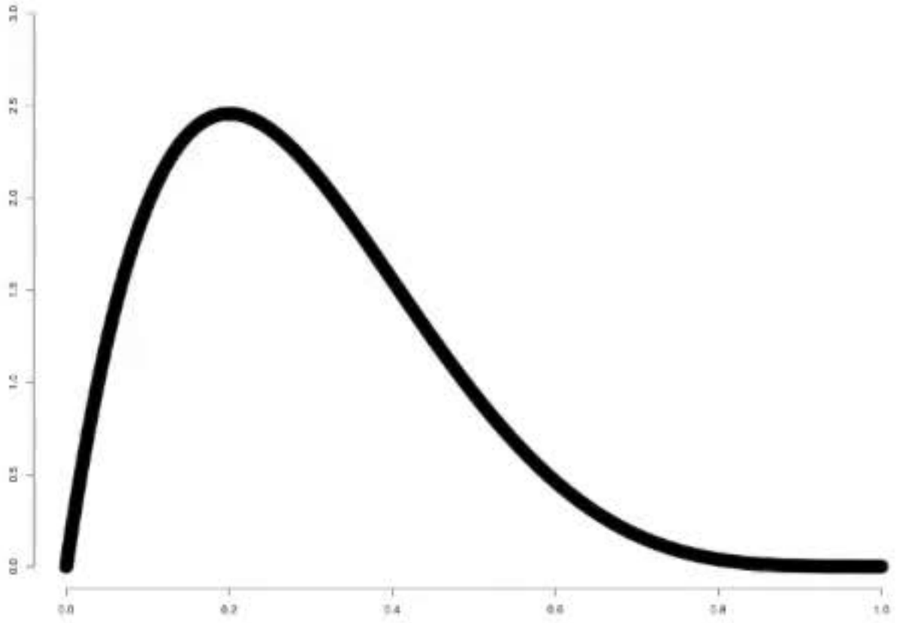


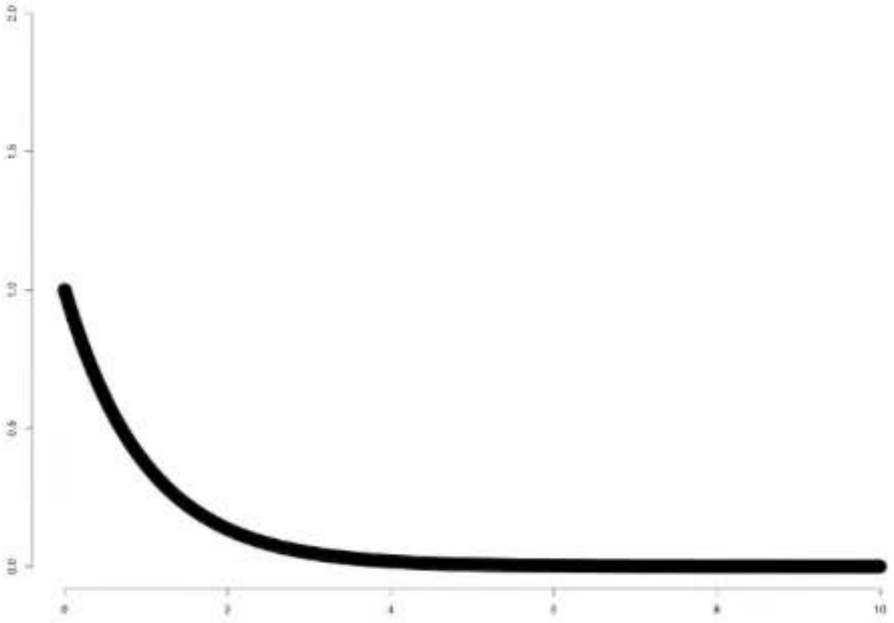


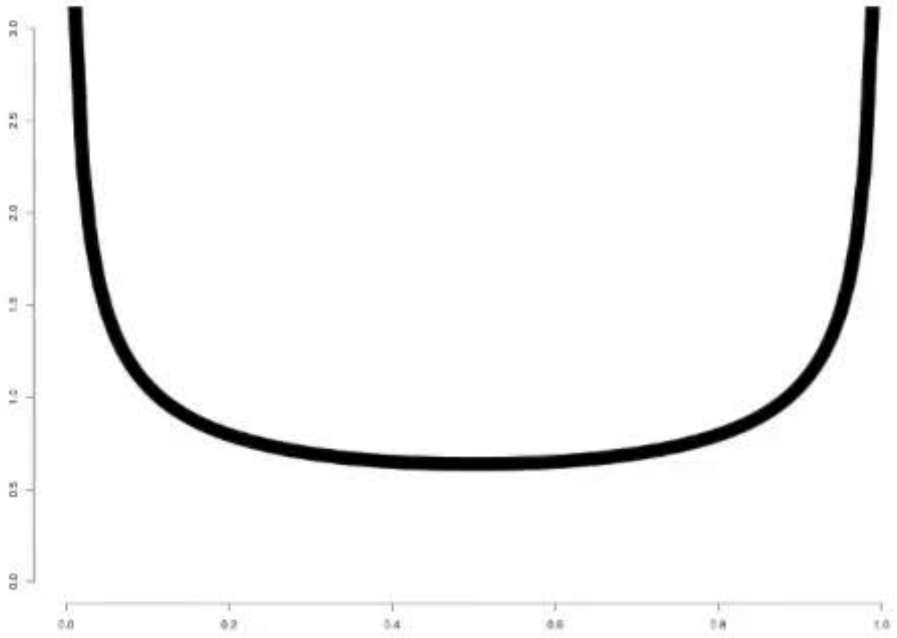
…are also normally distributed.

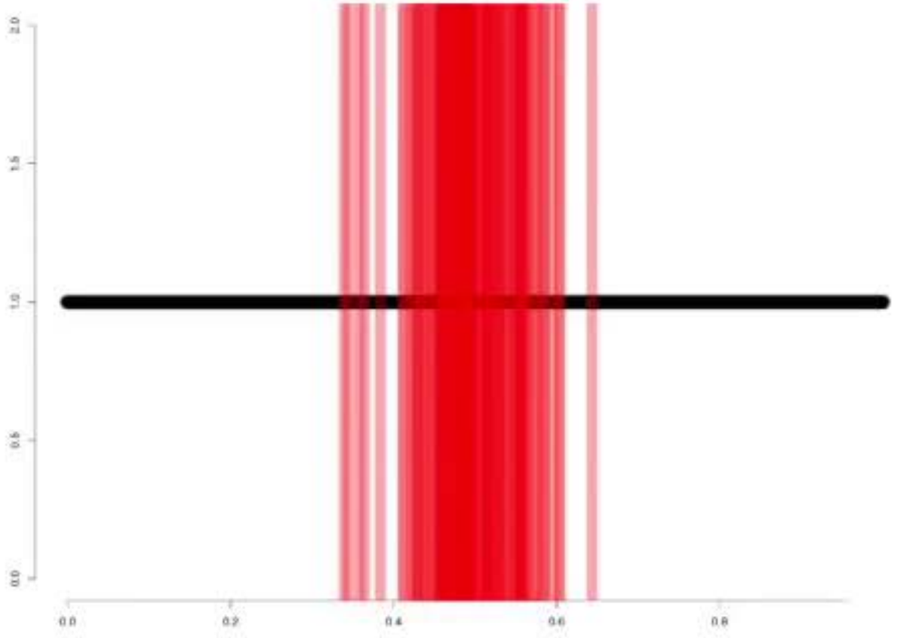


Well, it turns out that it doesn't matter what distribution you start with…

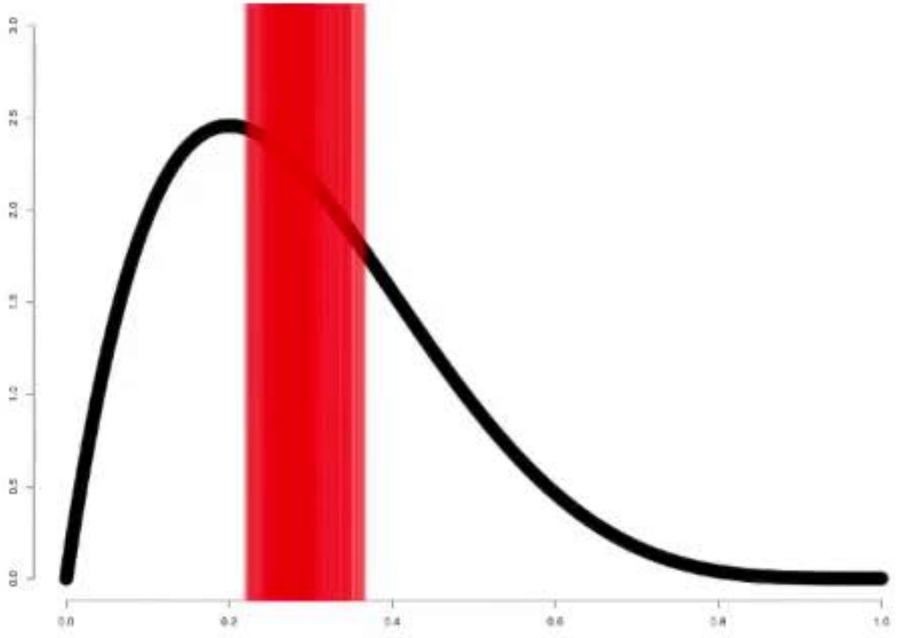


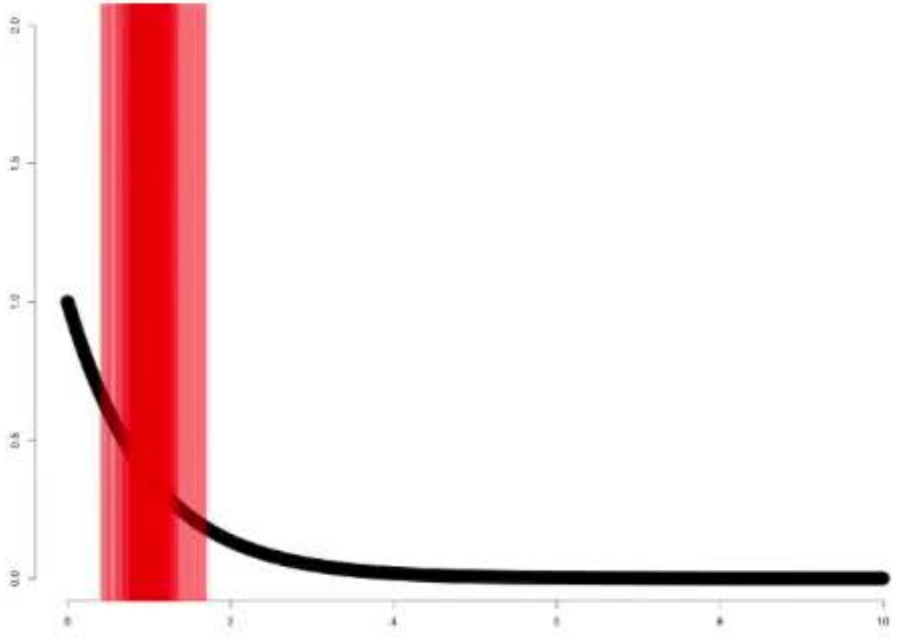


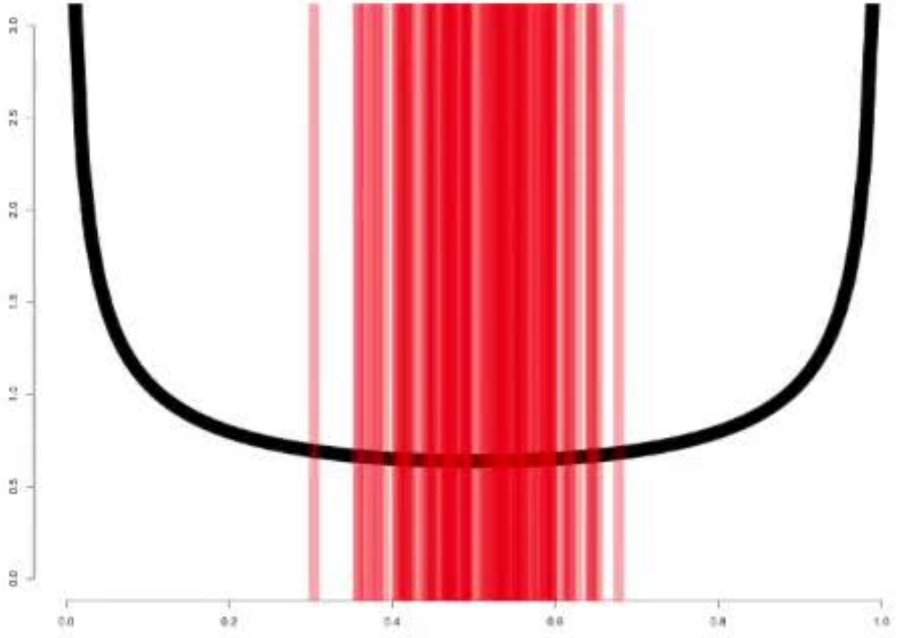


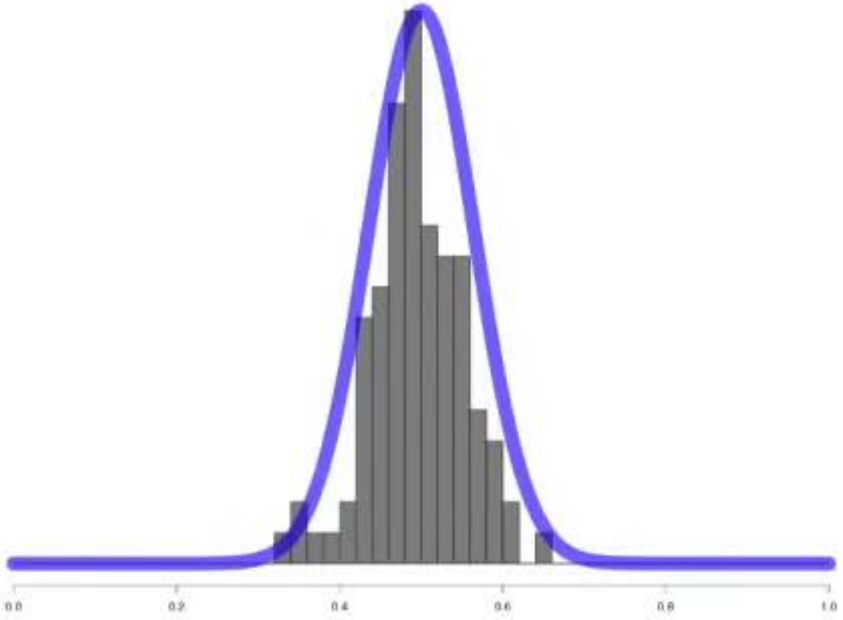


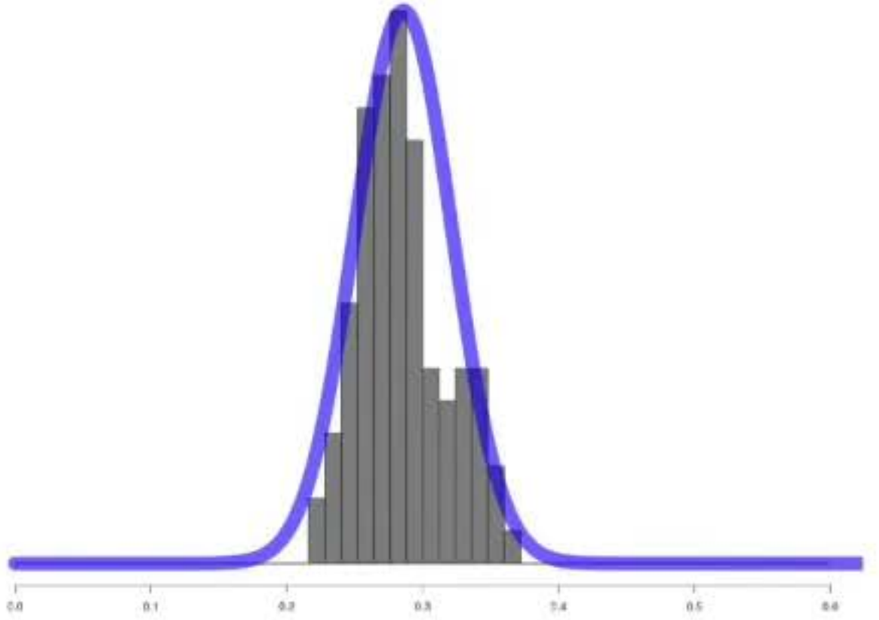
…if you collect samples from those distributions…



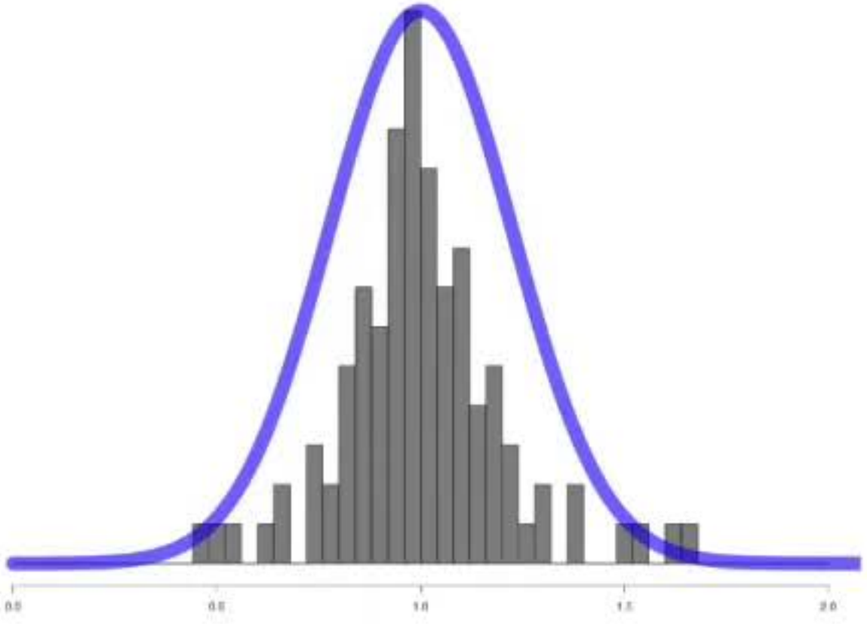


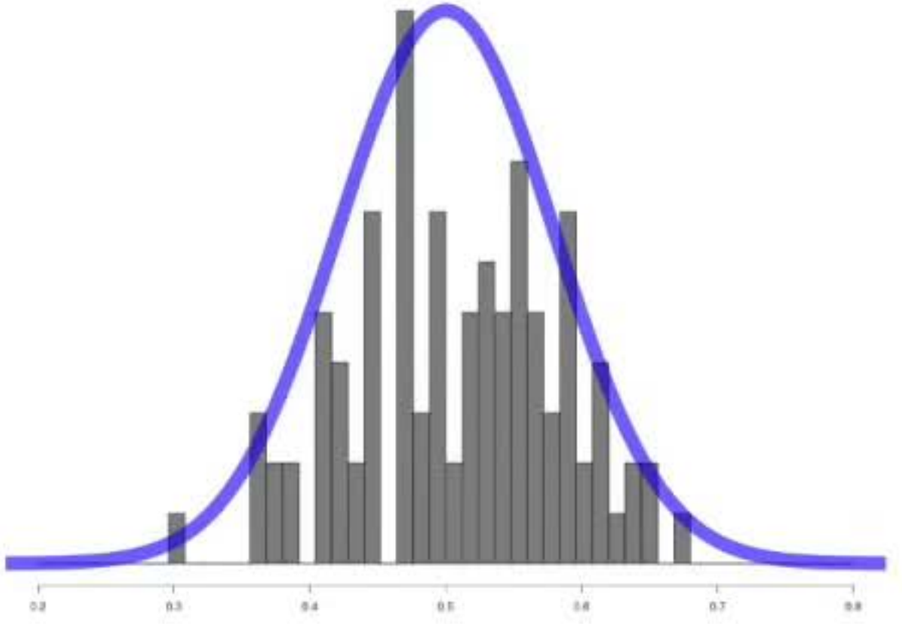






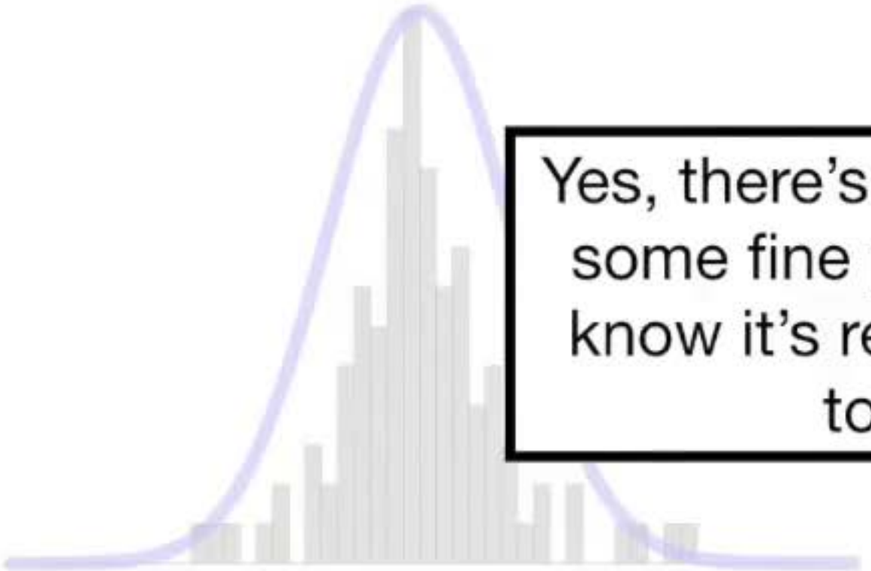
…then **the means will be normally distributed***.

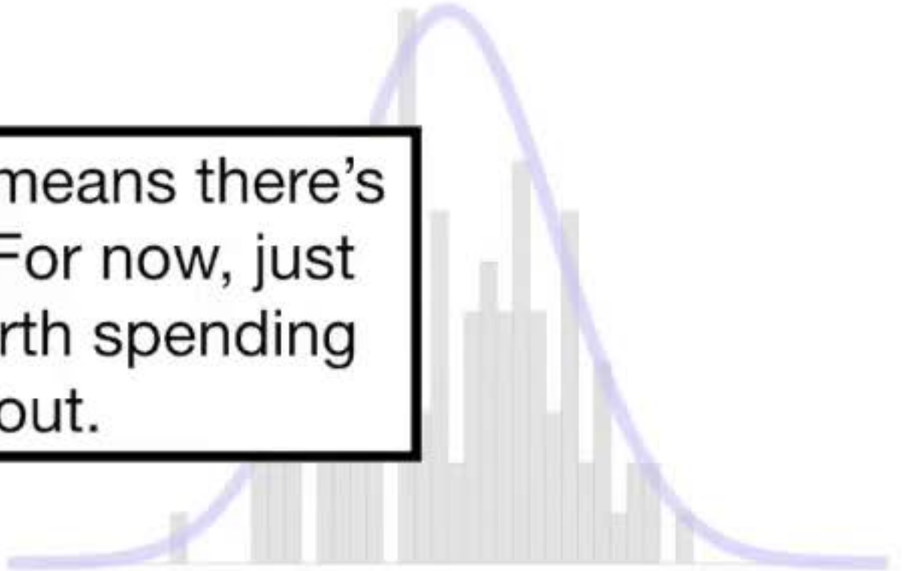




…then **the means will be normally distributed*.**

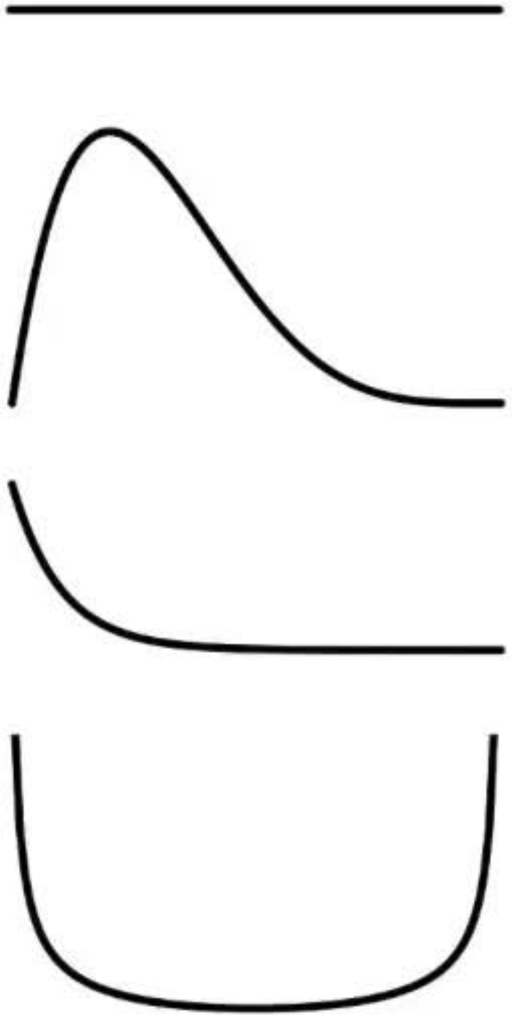
Yes, there's a little asterisk here that means there's some fine print that will come later. For now, just know it's really fine print and not worth spending too much time worrying about.





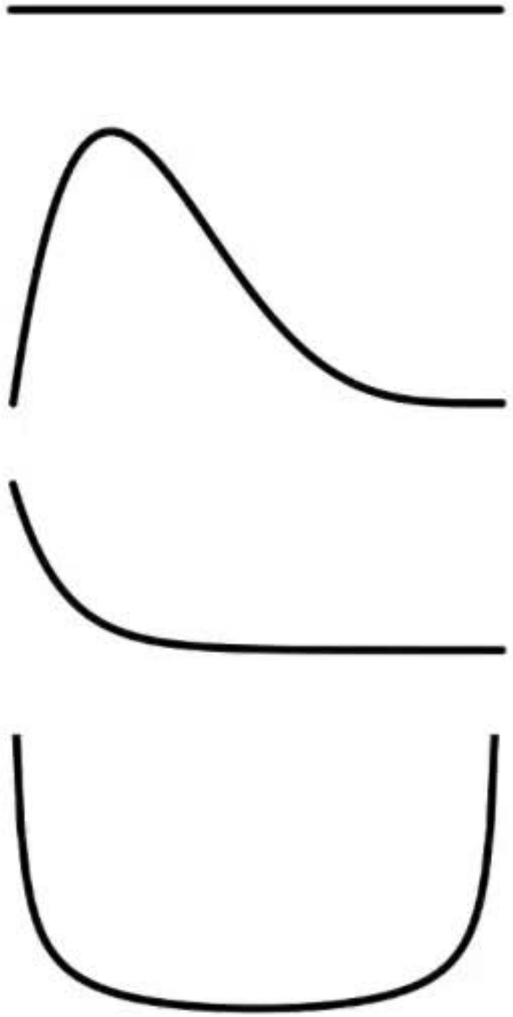
Cool!

But what are the practical implications of knowing that means are normally distributed?



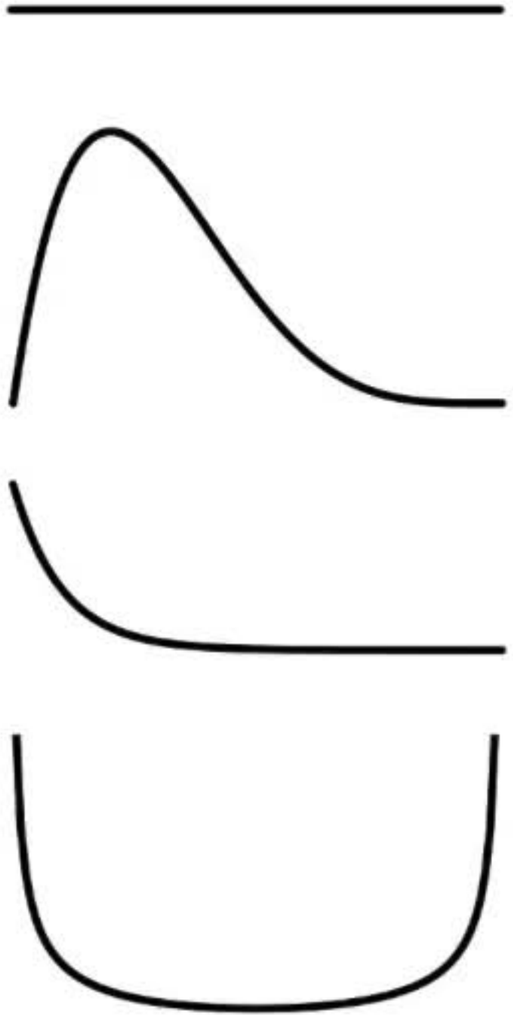
?

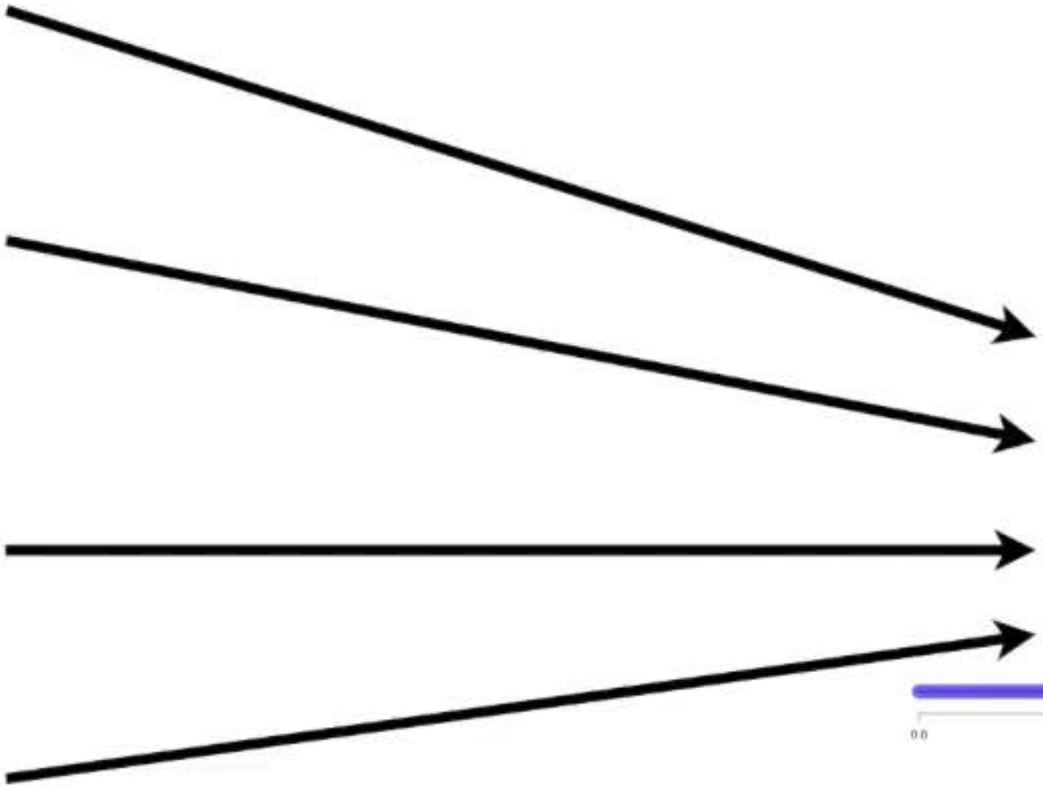
When we do an experiment, we don't always know what distribution our data comes from.



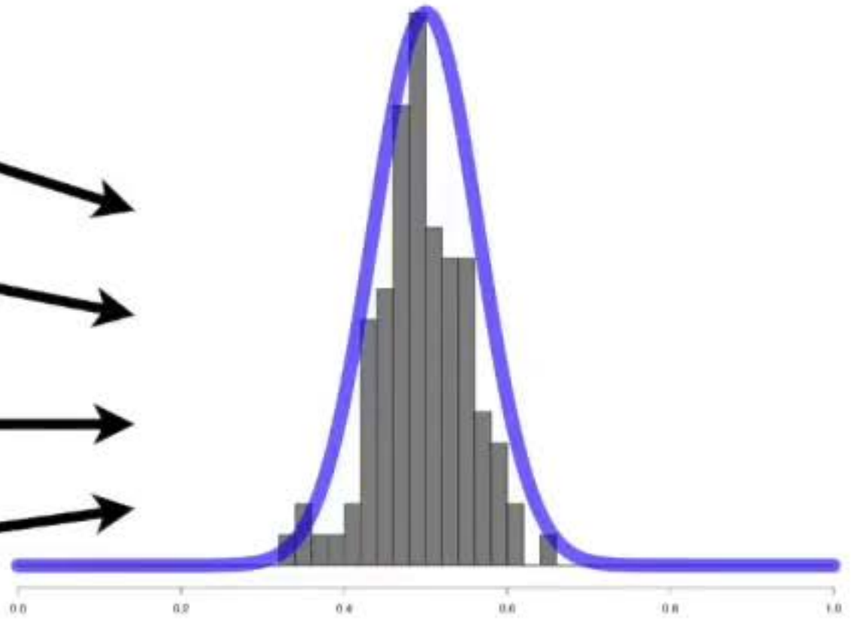
?

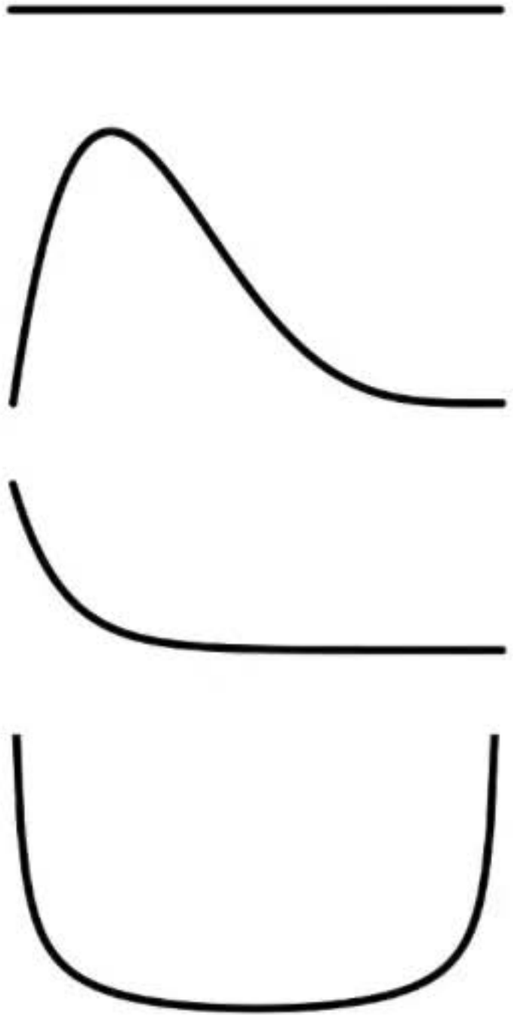
To this, **The Central Limit Theorem** says, **“Who Cares???”**

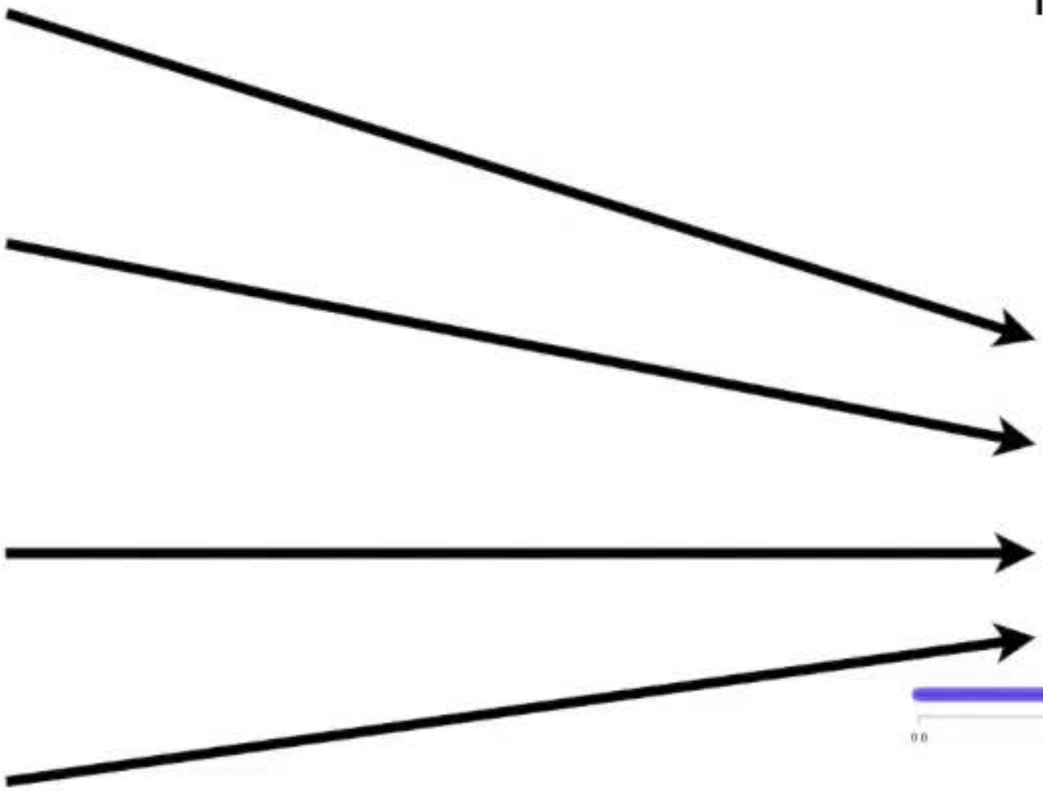




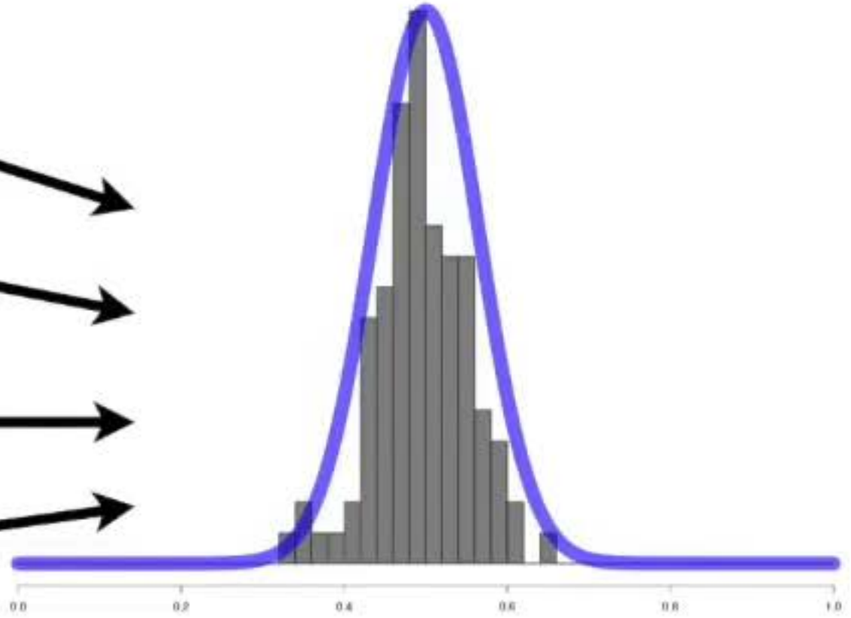
The sample means will be normally distributed.

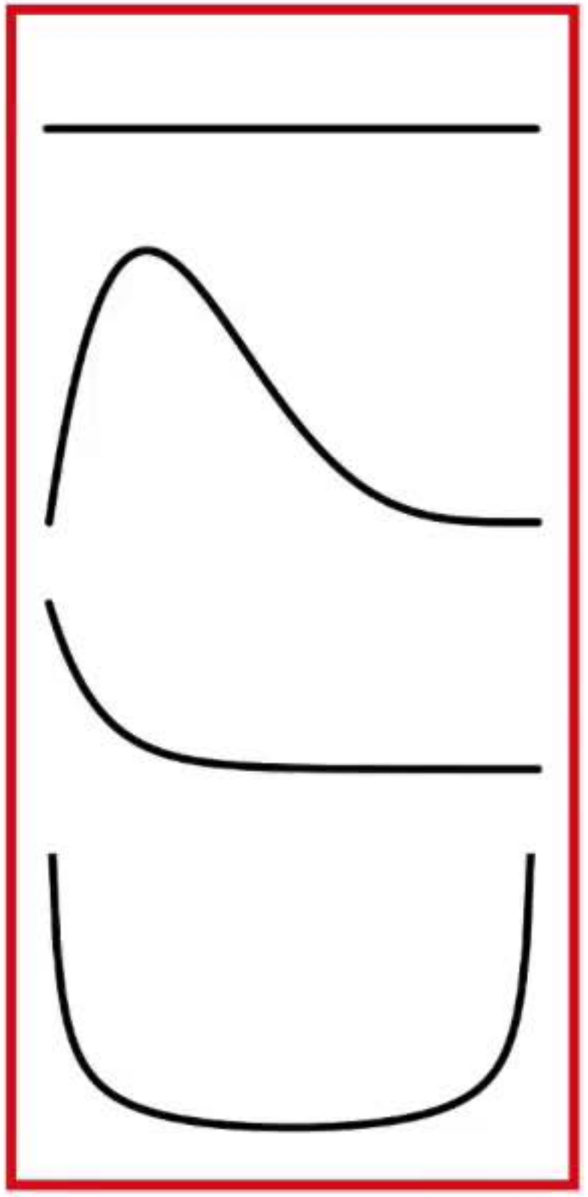




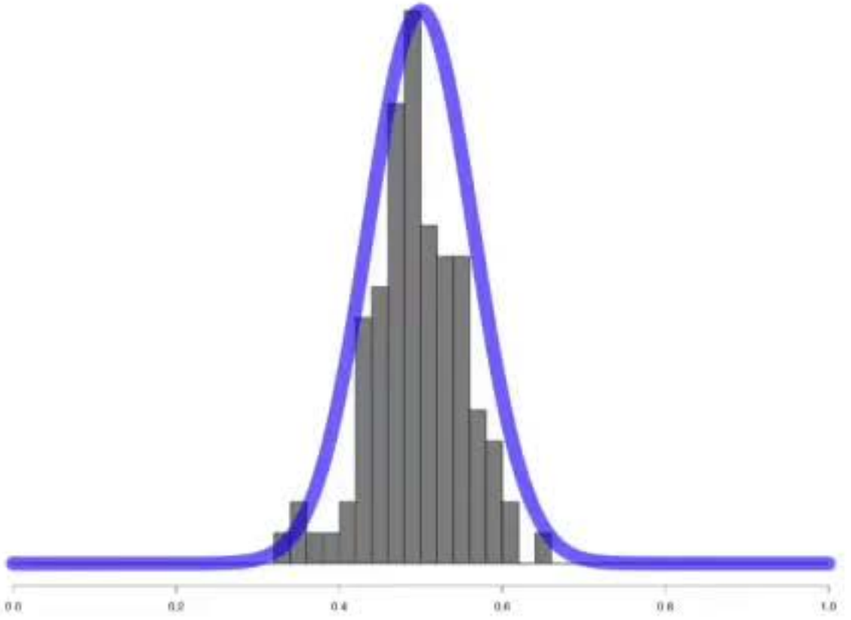


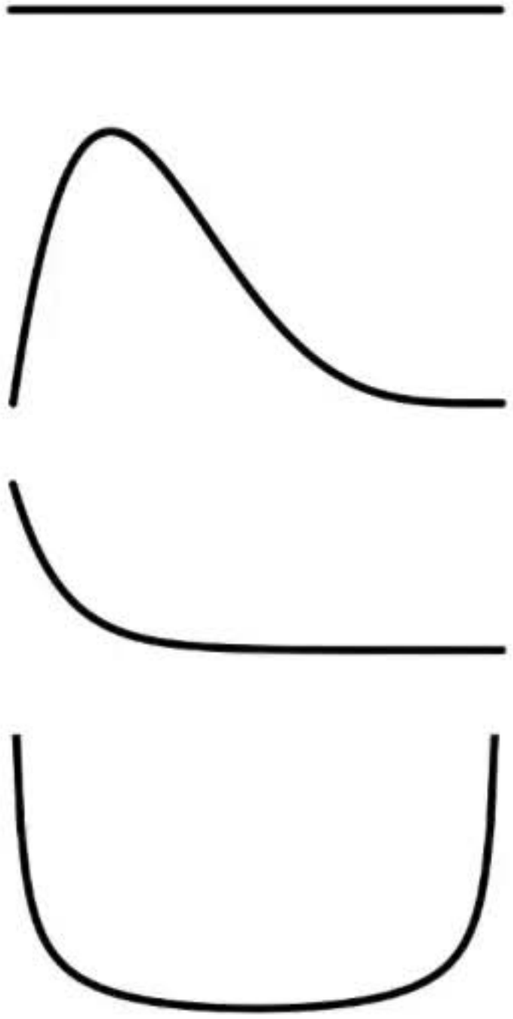
Because we know that the sample means are normally distributed…



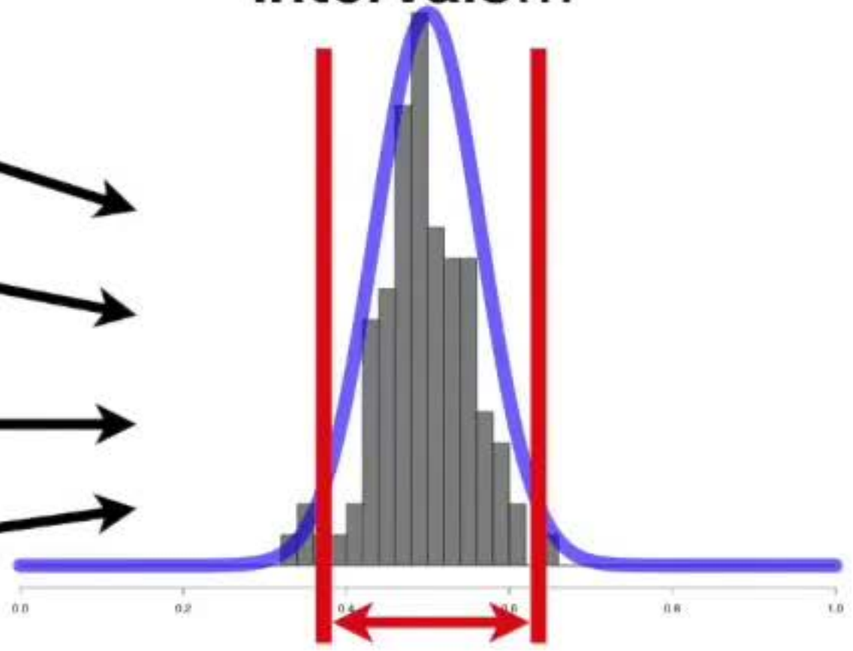


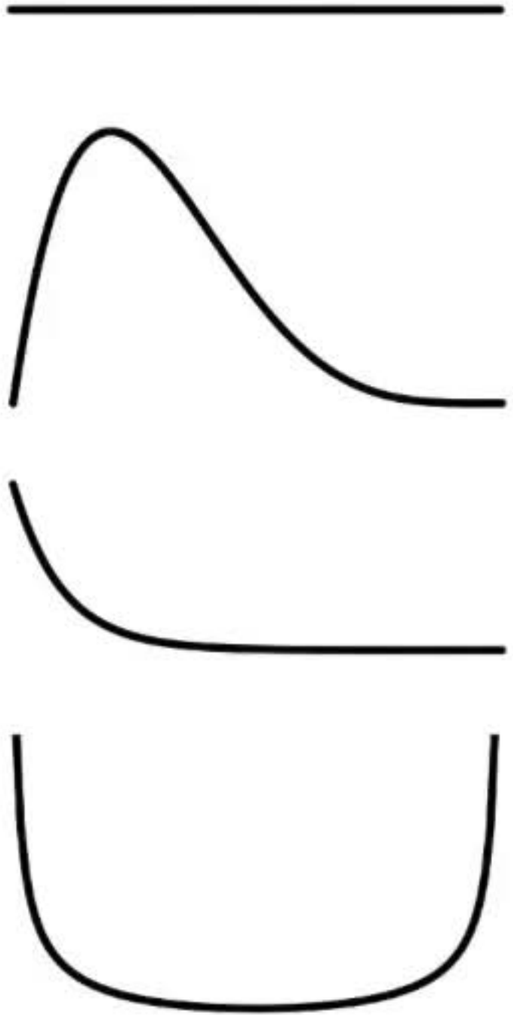
…we don’t need to worry too much about the distribution that the samples came from.



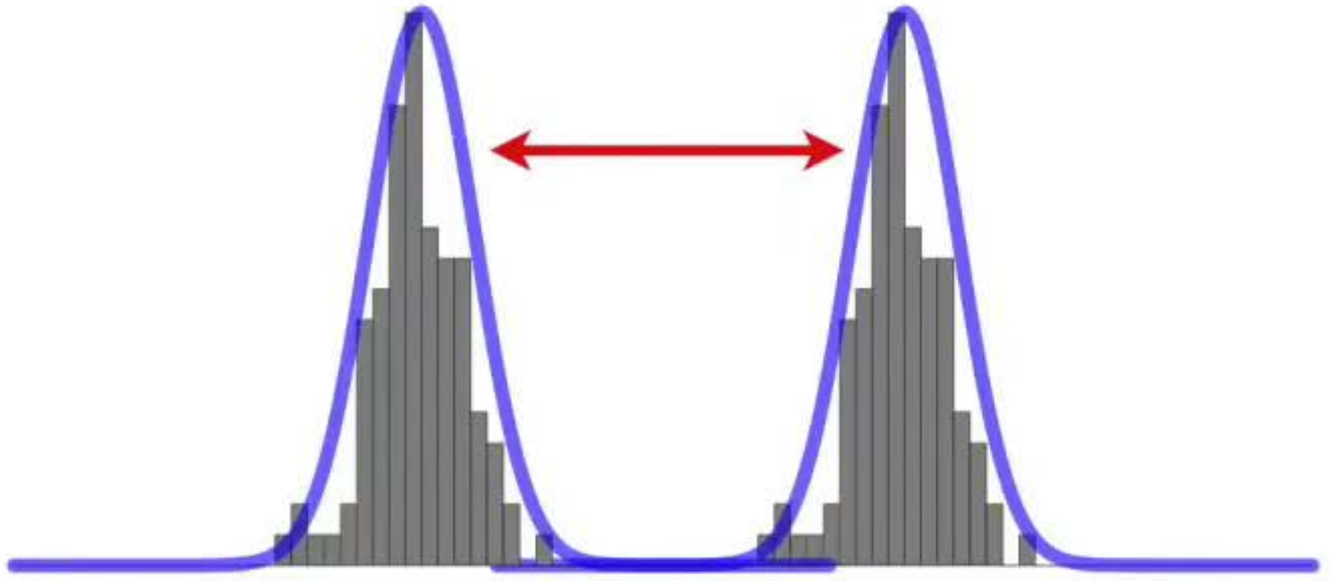


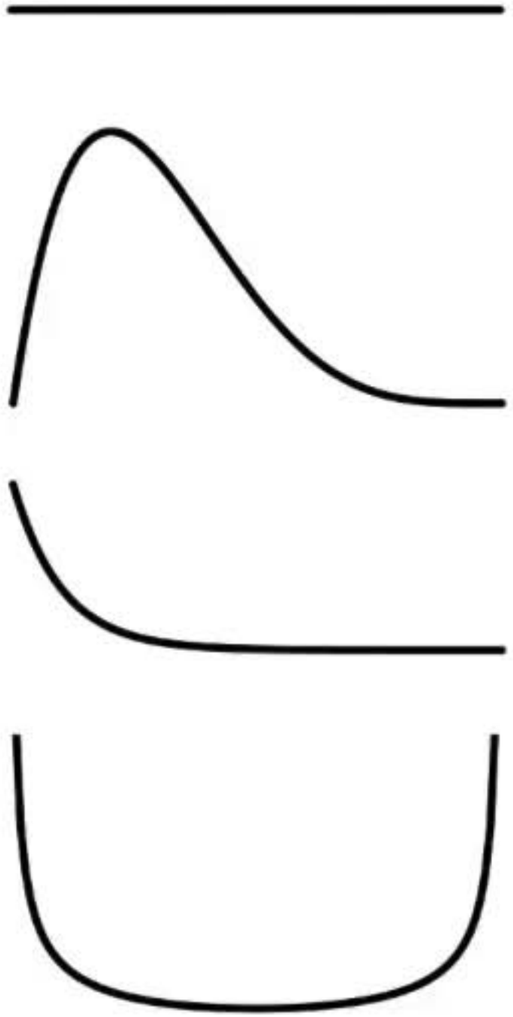
We can use the mean's normal distribution to make **confidence intervals**…



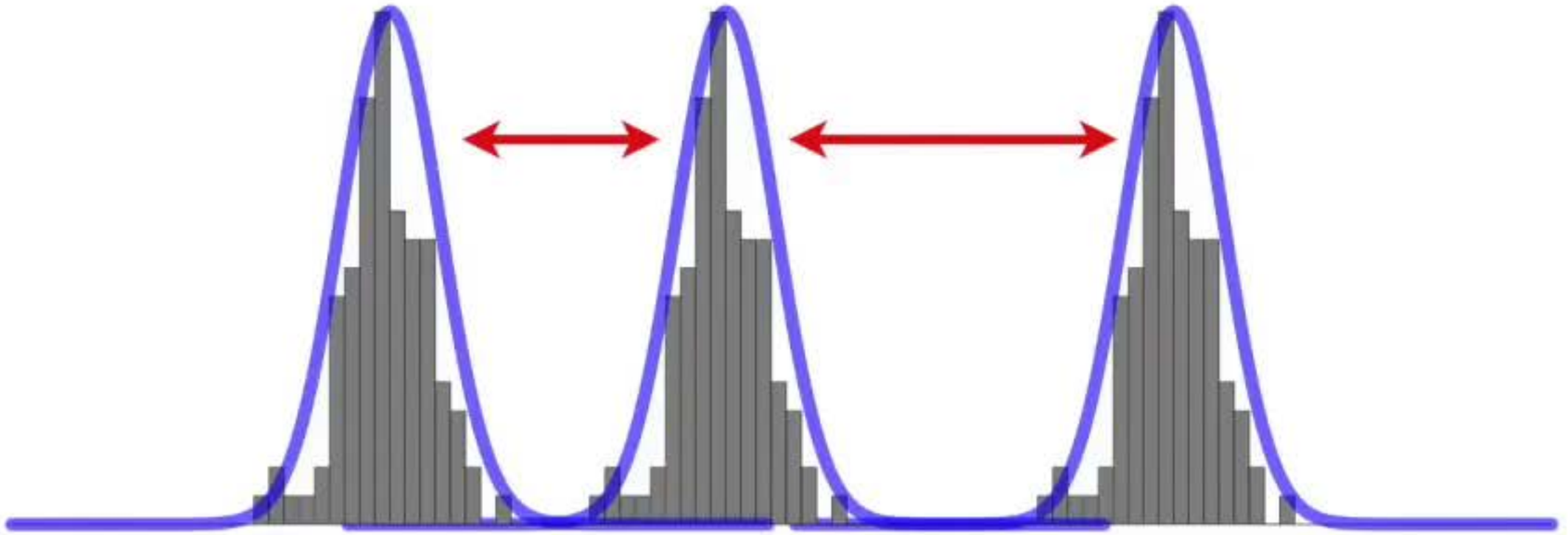


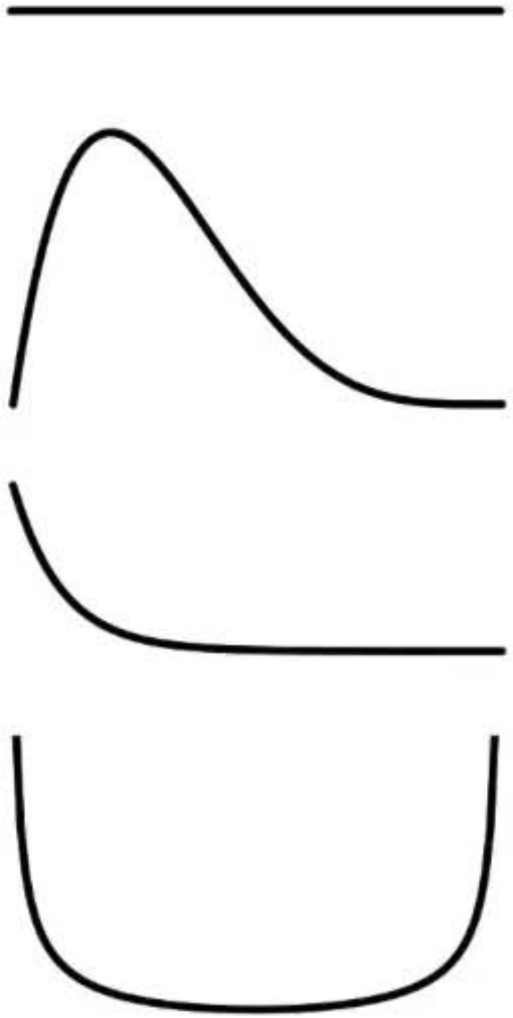
…do **t-tests**, where we ask if there is a difference between the means from two samples…



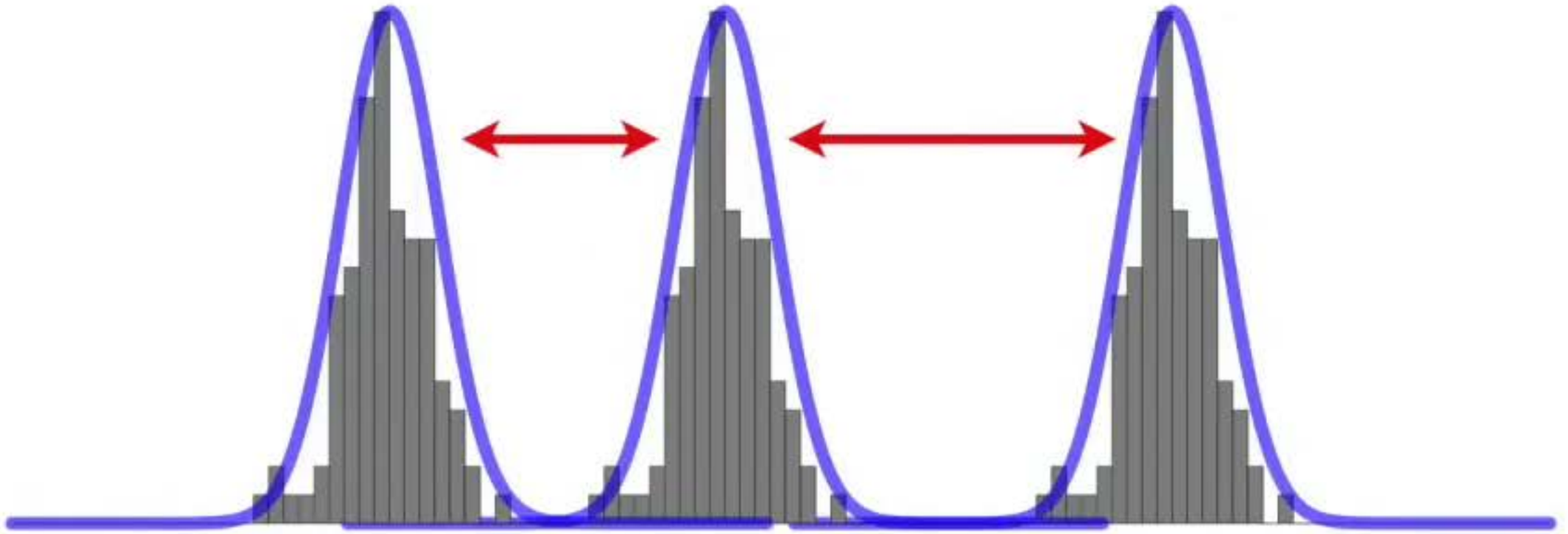


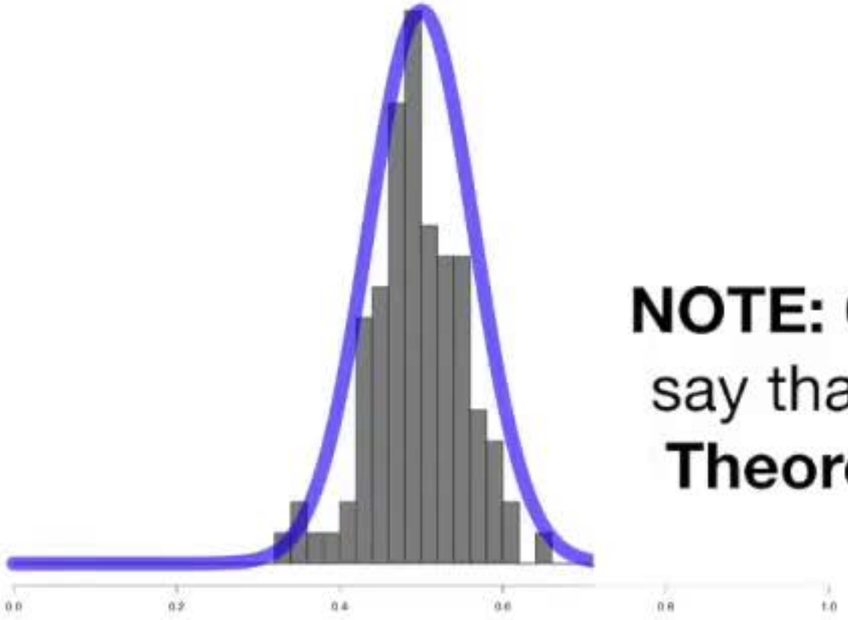
…and **ANOVA**, where we ask if there is a difference among the means from three or more samples…



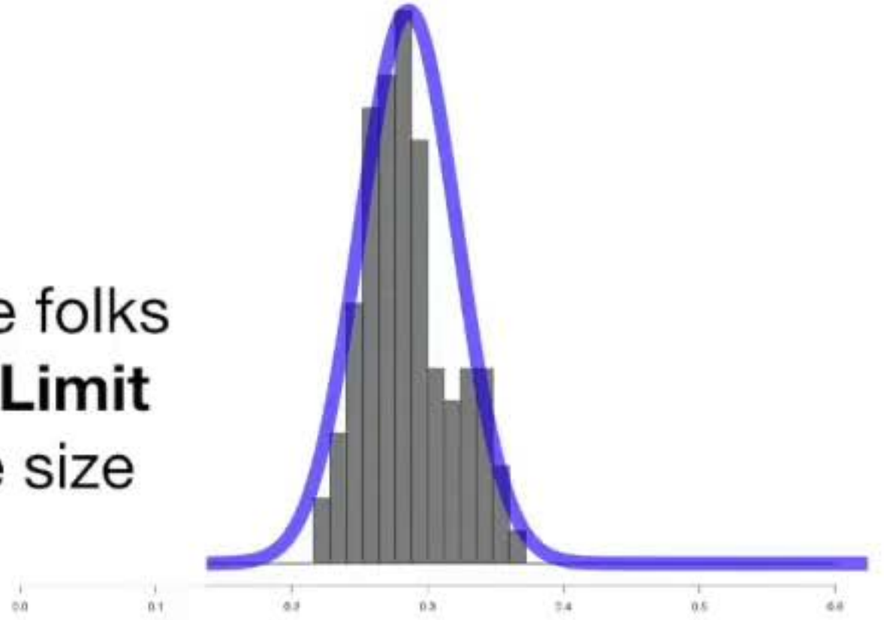


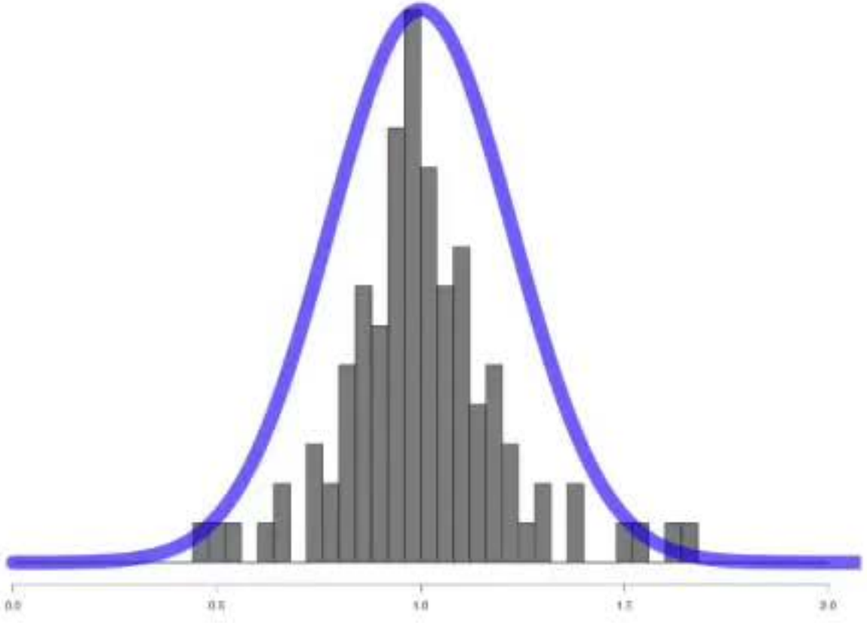
…and pretty much any statistical test
that uses the sample mean.

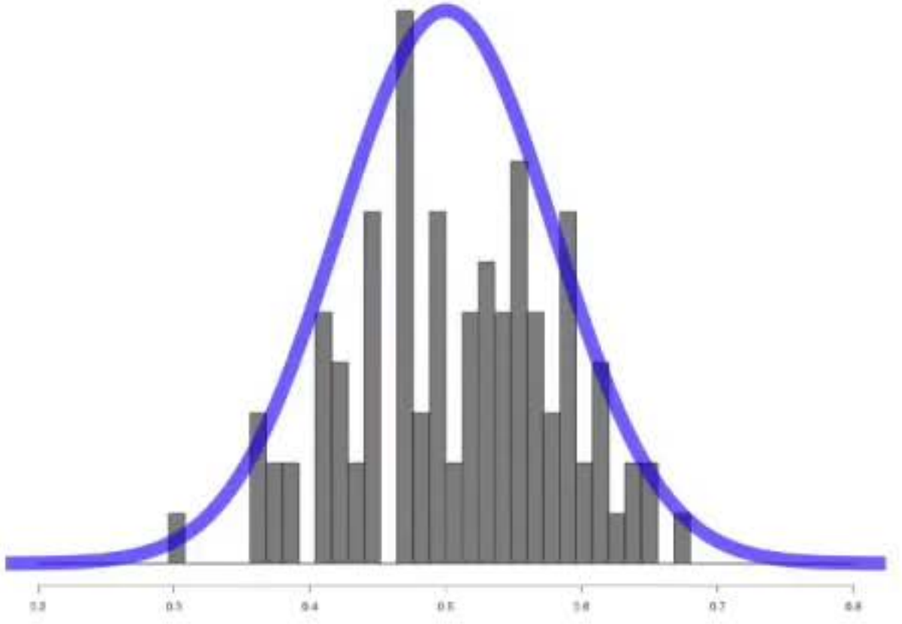


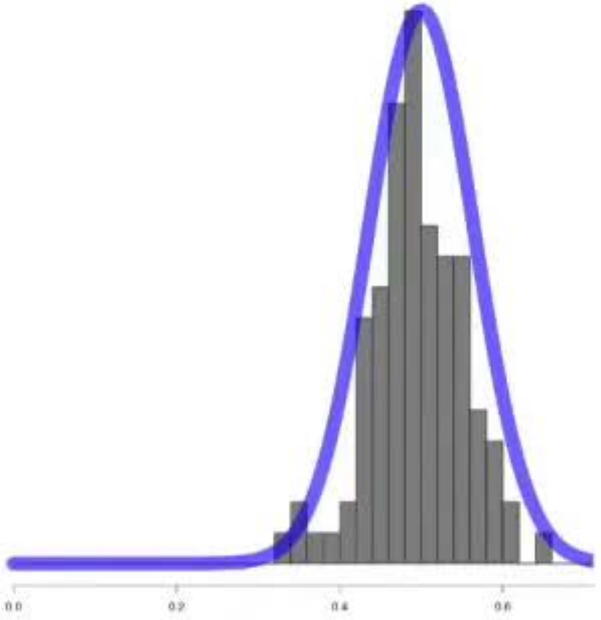


NOTE: Out there in the wild some folks say that in order for the **Central Limit Theorem** to be true, the sample size must be at least **30**.

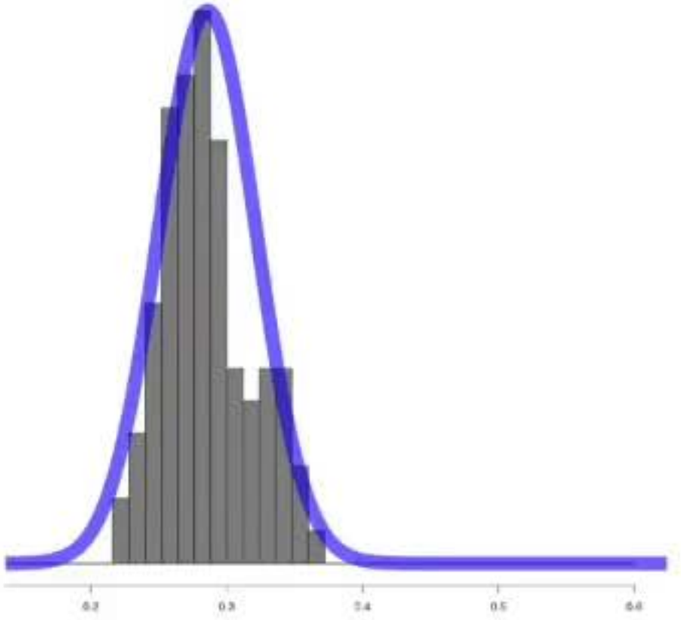




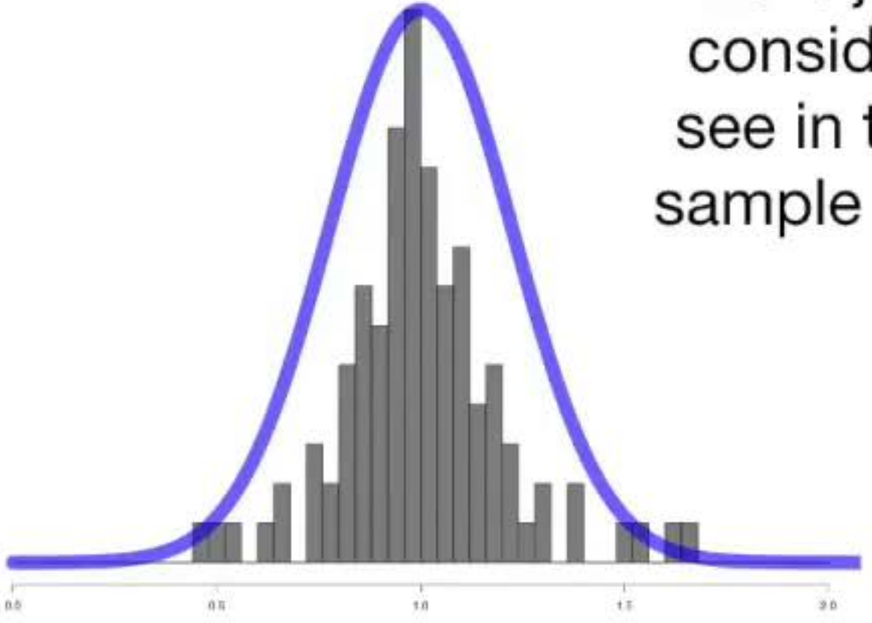


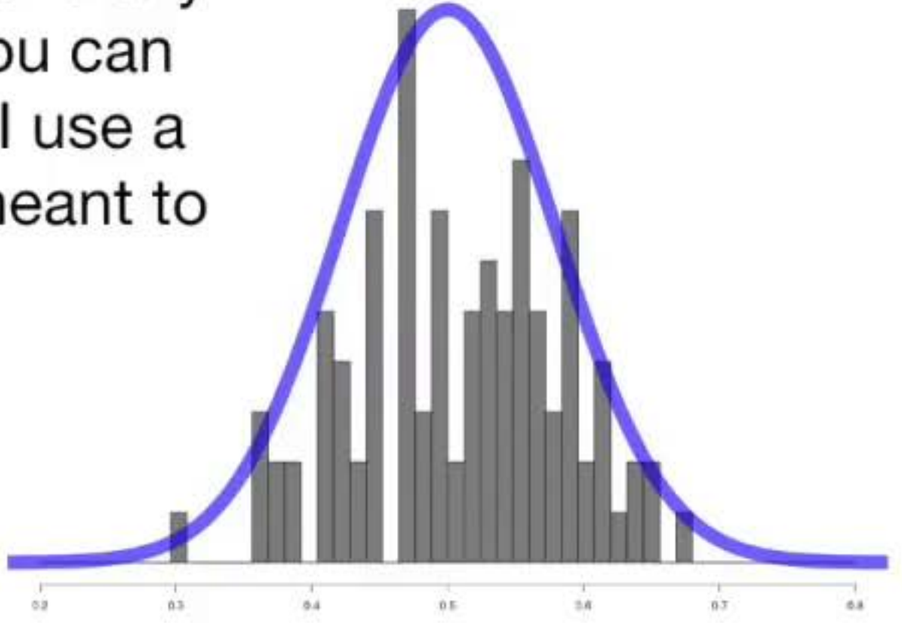


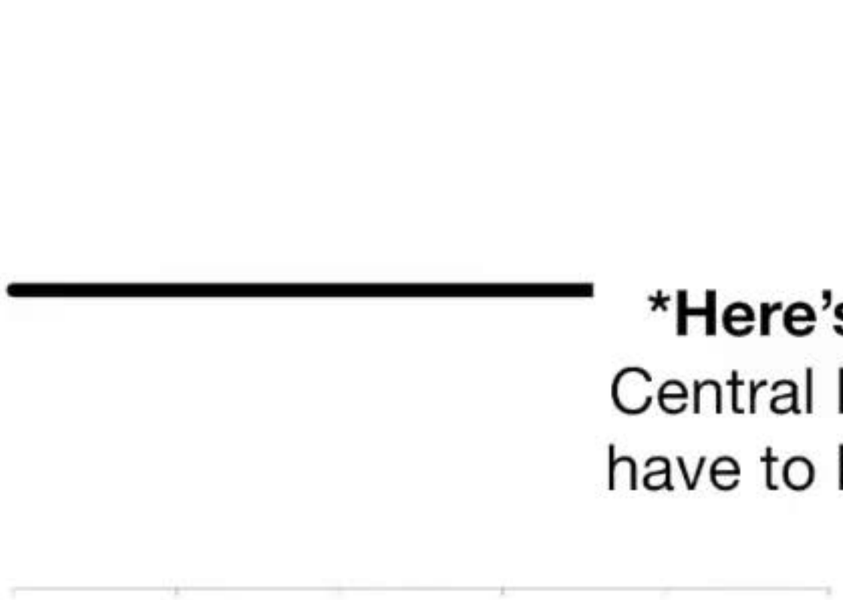
NOTE: Out there in the wild some folks say that in order for the **Central Limit Theorem** to be true, the sample size must be at least **30**.

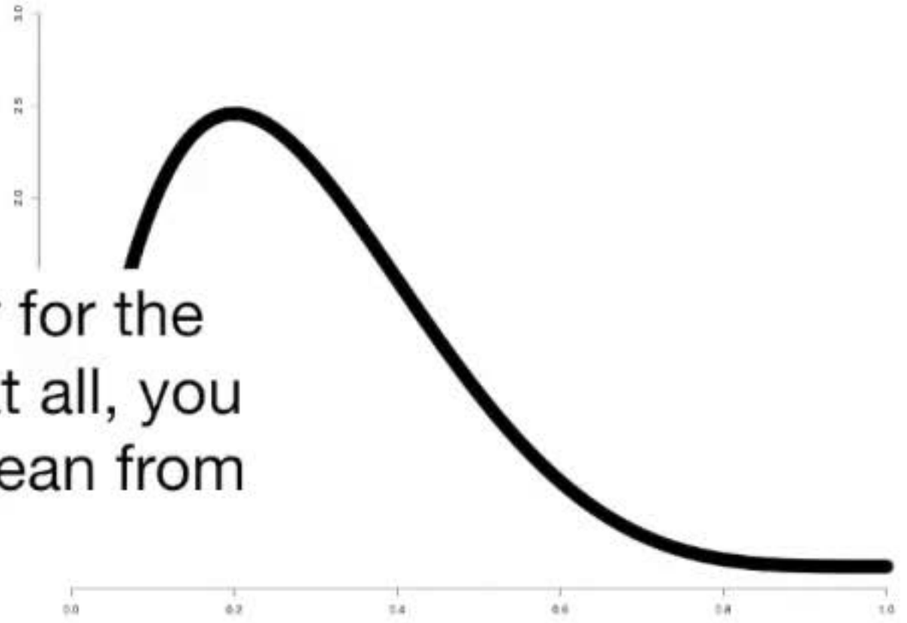


This is just a rule of thumb and generally considered safe. However, as you can see in the examples here where I use a sample size of **20**, the rule was meant to be broken.









***Here's the fine print:** In order for the Central Limit Theorem to work at all, you have to be able to calculate a mean from your sample.