Benford's Law

SING meeting 6.9.2010

Behram Mistree

Simple Question

What happens if we take first non-zero digits from a group of numbers?

Simple Question

What happens if we take first non-zero digits from a group of numbers?

0.323

1339.13

-553

22000

Simple Question

What happens if we take first non-zero digits from a group of numbers?

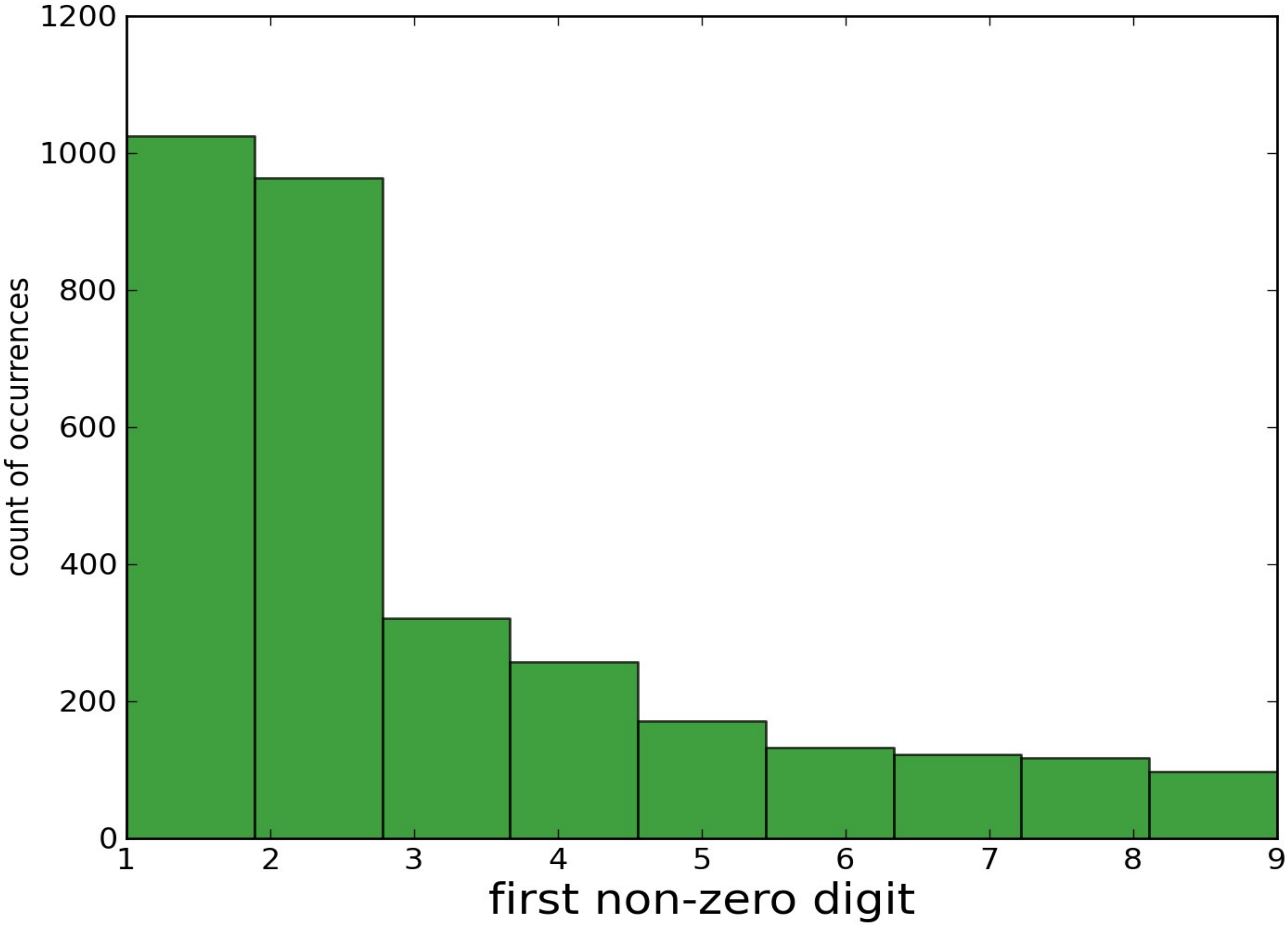
0.323

1339.13 → [3,1,5,2]

-553

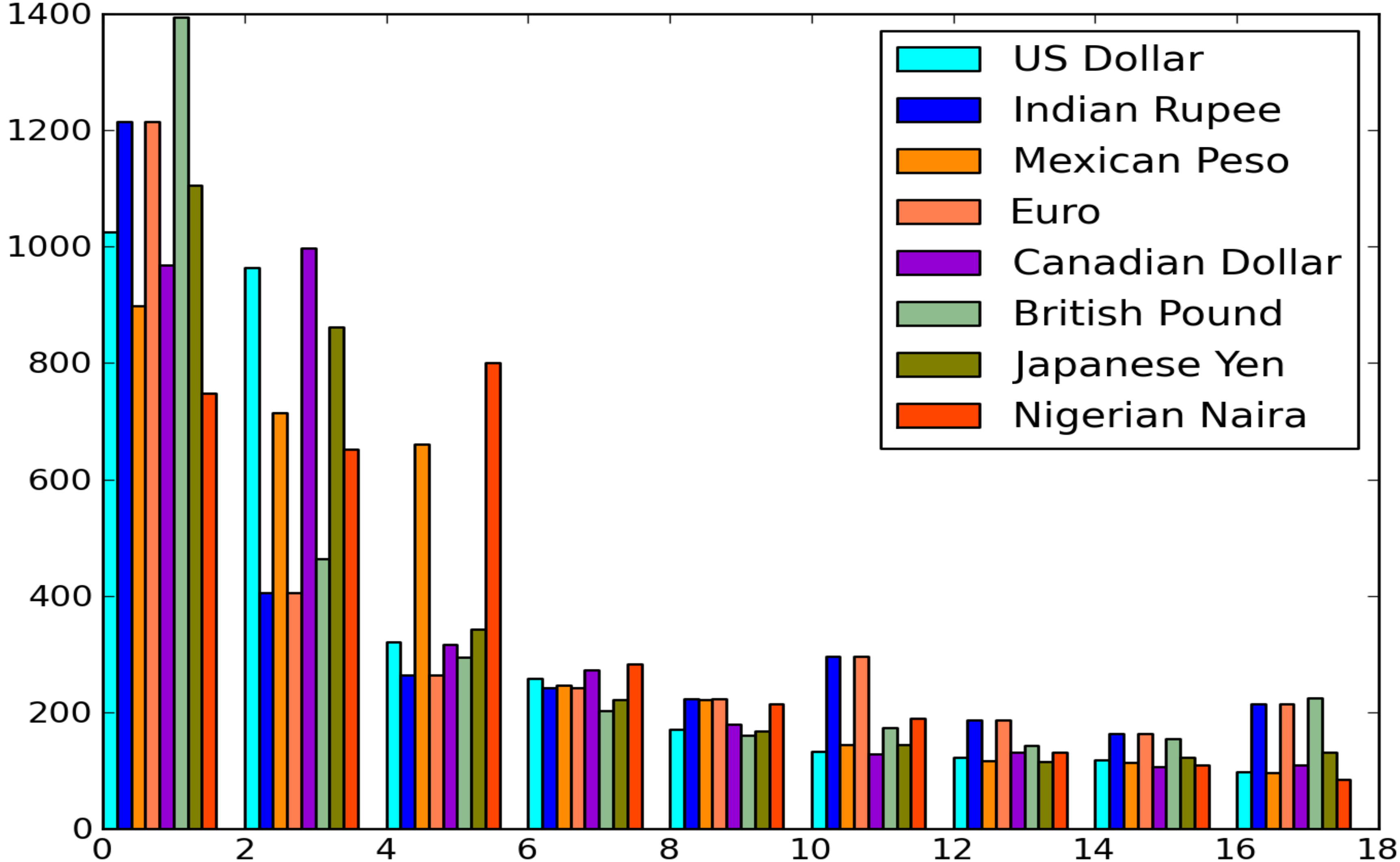
22000

First digit NYSE stock prices



What if that's just an artifact of US currency?

First digit NYSE prices different currencies



Maybe it's just a weird artifact of stock data?

Maybe it's just a weird artifact of the data?

- Physical constants

Maybe it's just a weird artifact of stock data?

- Physical constants
- Census data

Maybe it's just a weird artifact of stock data?

- Physical constants
- Census data
- Numbers in *The Farmer's Almanac*

Maybe it's just a weird artifact of stock data?

- Physical constants
- Census data
- Numbers in *The Farmer's Almanac*
- Lengths of the world's rivers

Maybe it's just a weird artifact of stock data?

- Physical constants
- Census data
- Numbers in *The Farmer's Almanac*
- Lengths of the world's rivers

All have pronouncedly more first non-zero digit 1's than any other value. Specifically:

$p(1) = .301$

(even when multiplied by any constants)

Rest of this talk: Benfordian Scale invariance

- Computational test for scale invariance in Benford's Law
- Math-ish test for scale invariance in Benford's Law
- Explanation of what's actually happening

References

Talk heavily informed from the following 5 references:

- *The Scientist and Engineer's Guide to Digital Signal Processing,* by Steven W. Smith
- “Looking out for number one” in *Plus Magazine*, by Jon Walthoe, Robert Hunt and Mike Pearson
- “A statistical derivation of the significant-digit law” in *Statistical Science*, by Theodore Hill. 1996
- EE261 Course Notes.
- Wikipedia

References

Talk heavily informed from the following 5 references:

- *The Scientist and Engineer's Guide to Digital Signal Processing,* by Steven W. Smith
- “Looking out for number one” in *Plus Magazine*, by Jon Walthoe, Robert Hunt and Mike Pearson
- “A statistical derivation of the significant-digit law” in *Statistical Science*, by Theodore Hill. 1996
- EE261 Course Notes.
- Wikipedia

Test a distribution's first digit scale-invariance (Computational)

Computationally, what would you do?

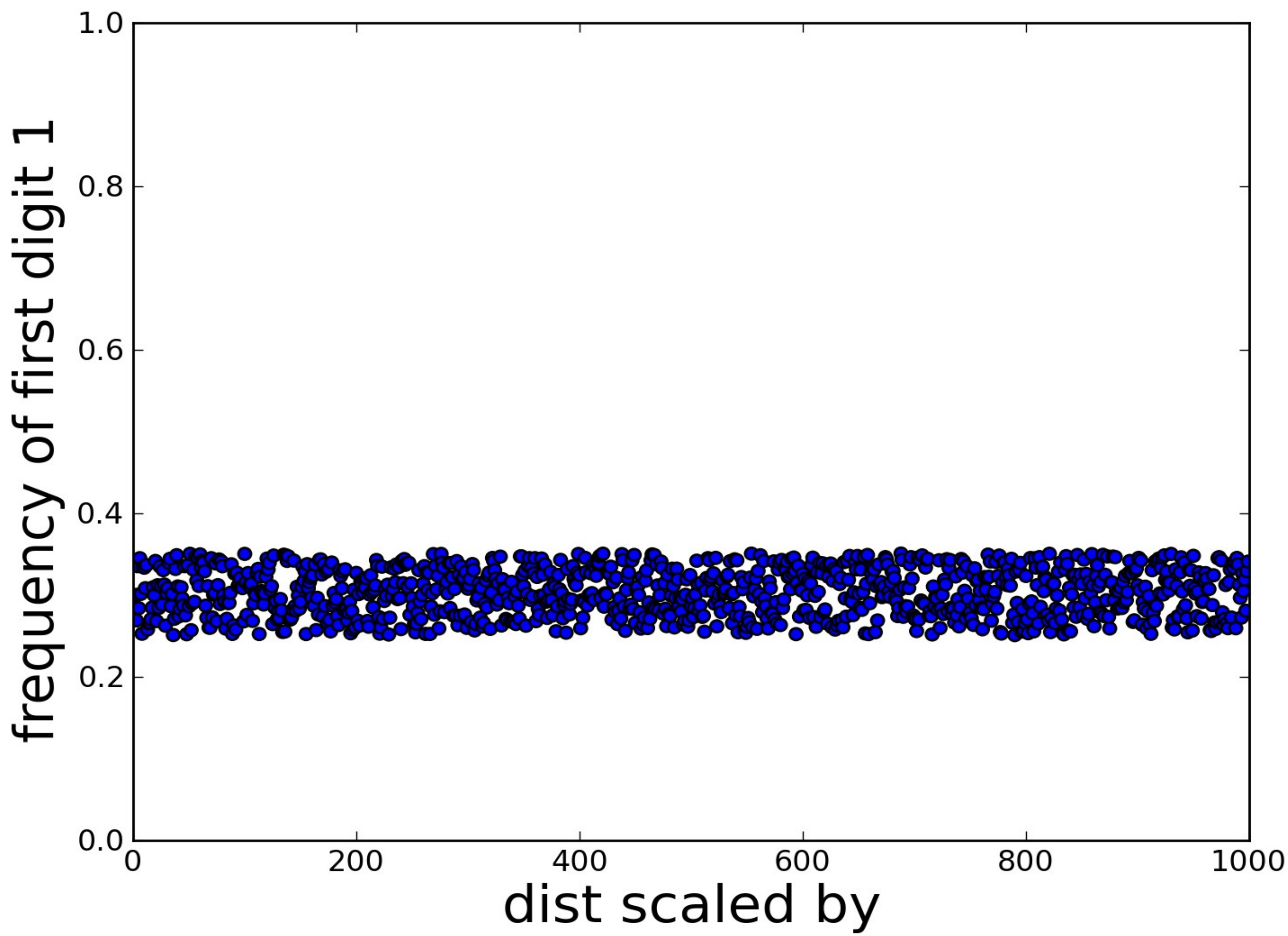
Test a distribution's first digit scale-invariance (Computational)

- Draw from distribution 1000 times. Count how many leading-digit 1's.
- Multiply all numbers in distribution by a constant and draw again.

Computational algorithm

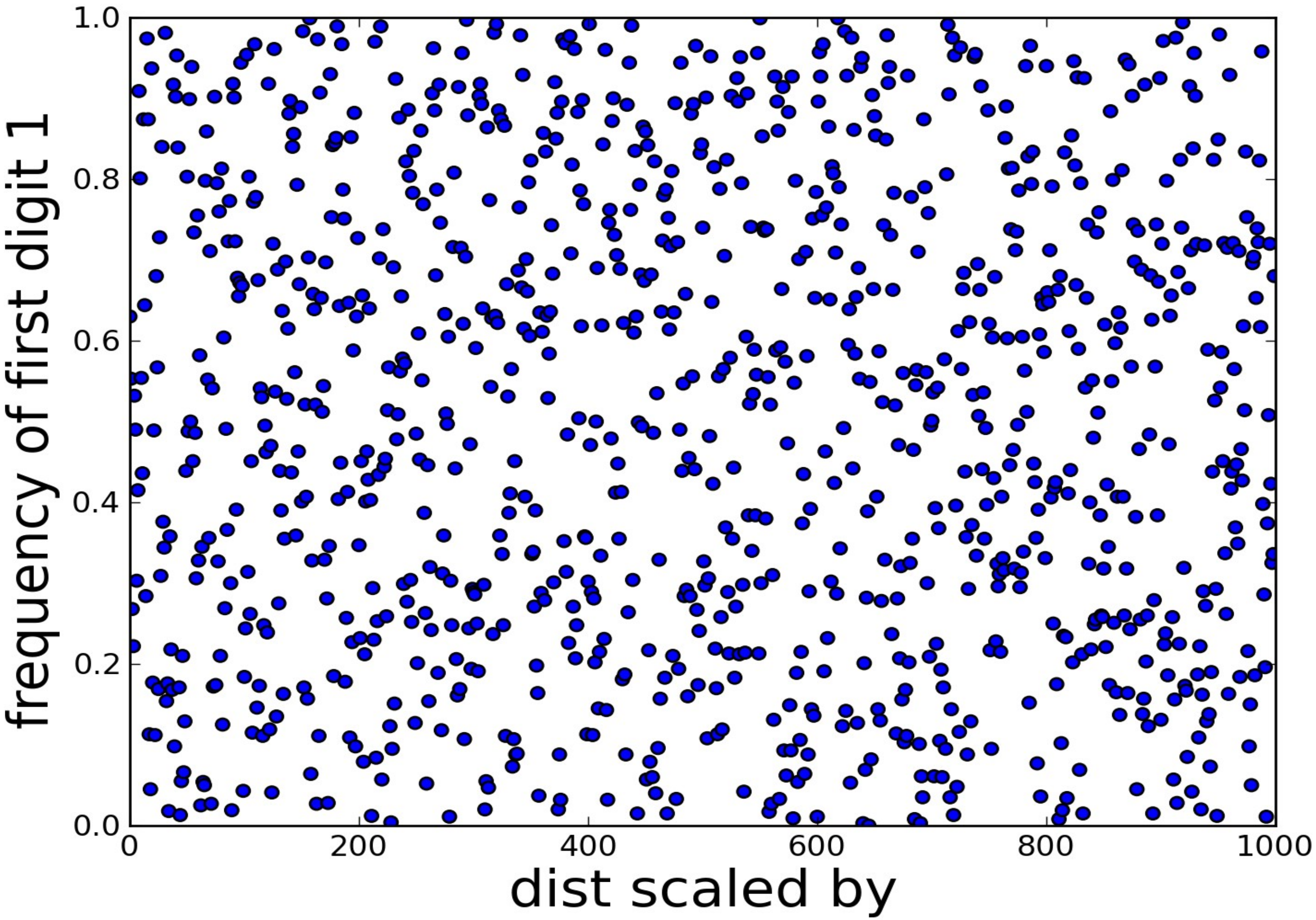
```
originalData = getData("filename.csv")
results= [];
for s = 1:.01:1000
   testData = originalData .* s;
   results.append(fracFirstDigitOnes(testData));

plot(results)
```

Benfordian!



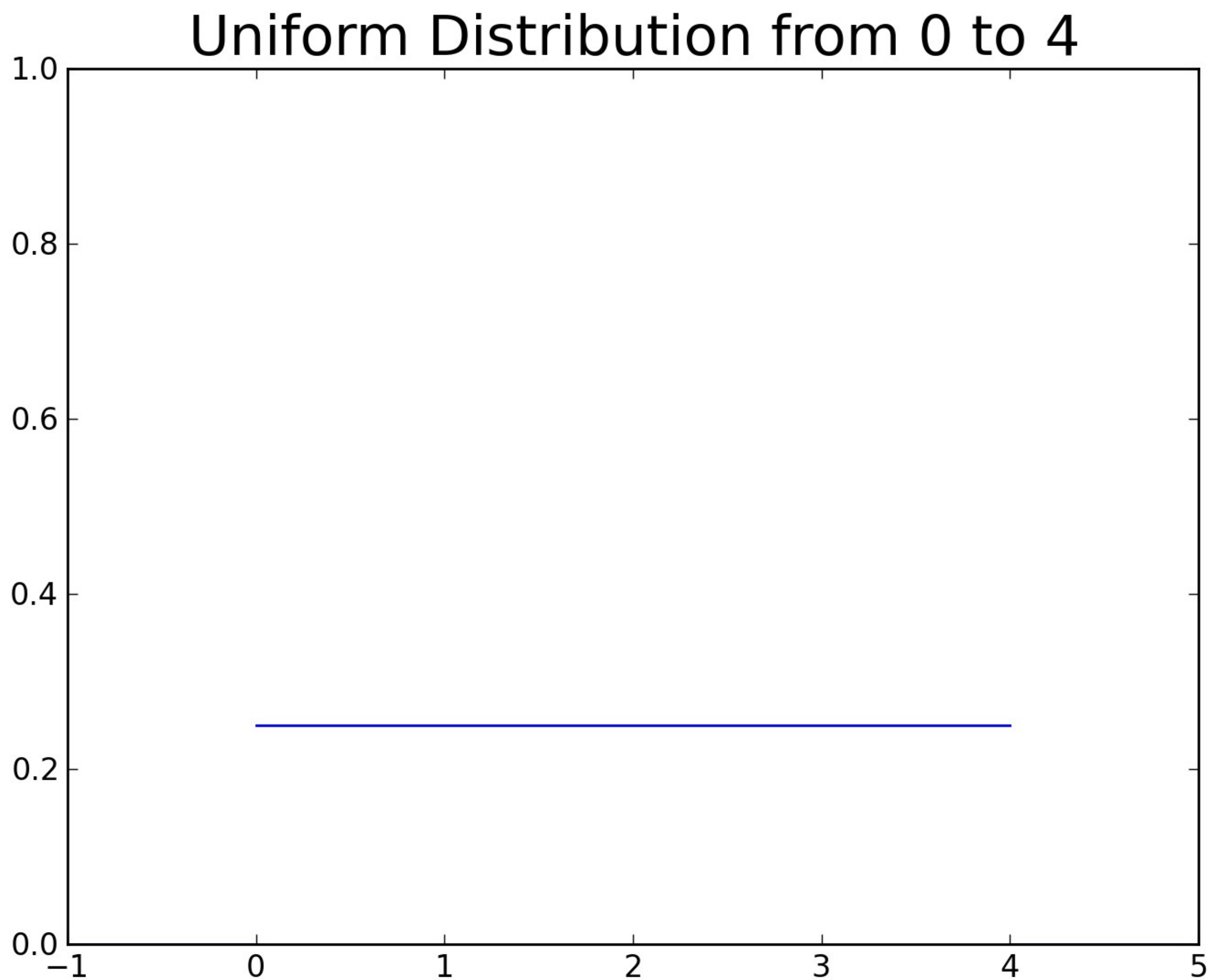
frequency of first digit 1
1.0
0.8
0.6
0.4
0.2
0.0
0
200
400
600
800
1000
dist scaled by

Non-Benfordian!

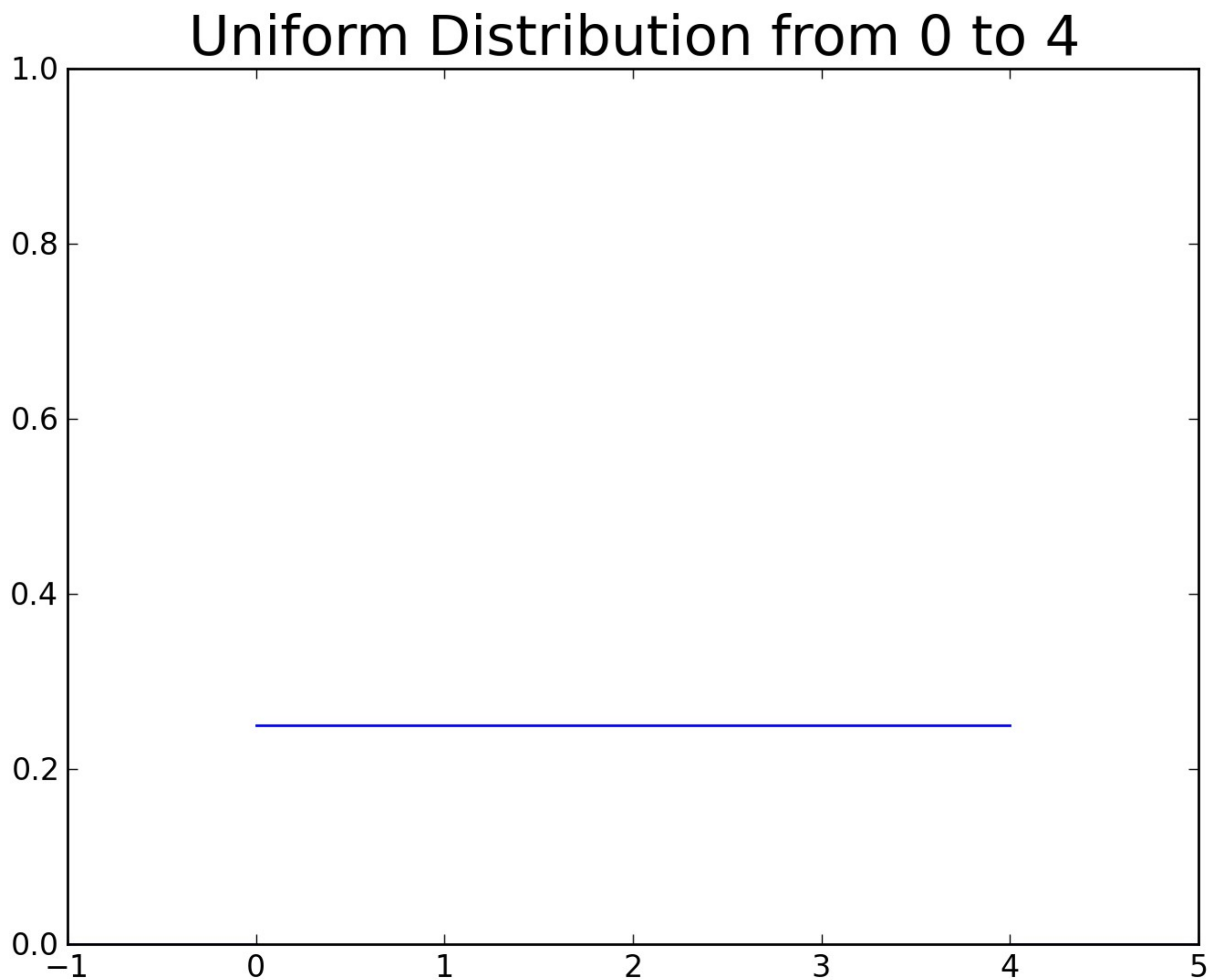


Let's be more math-y

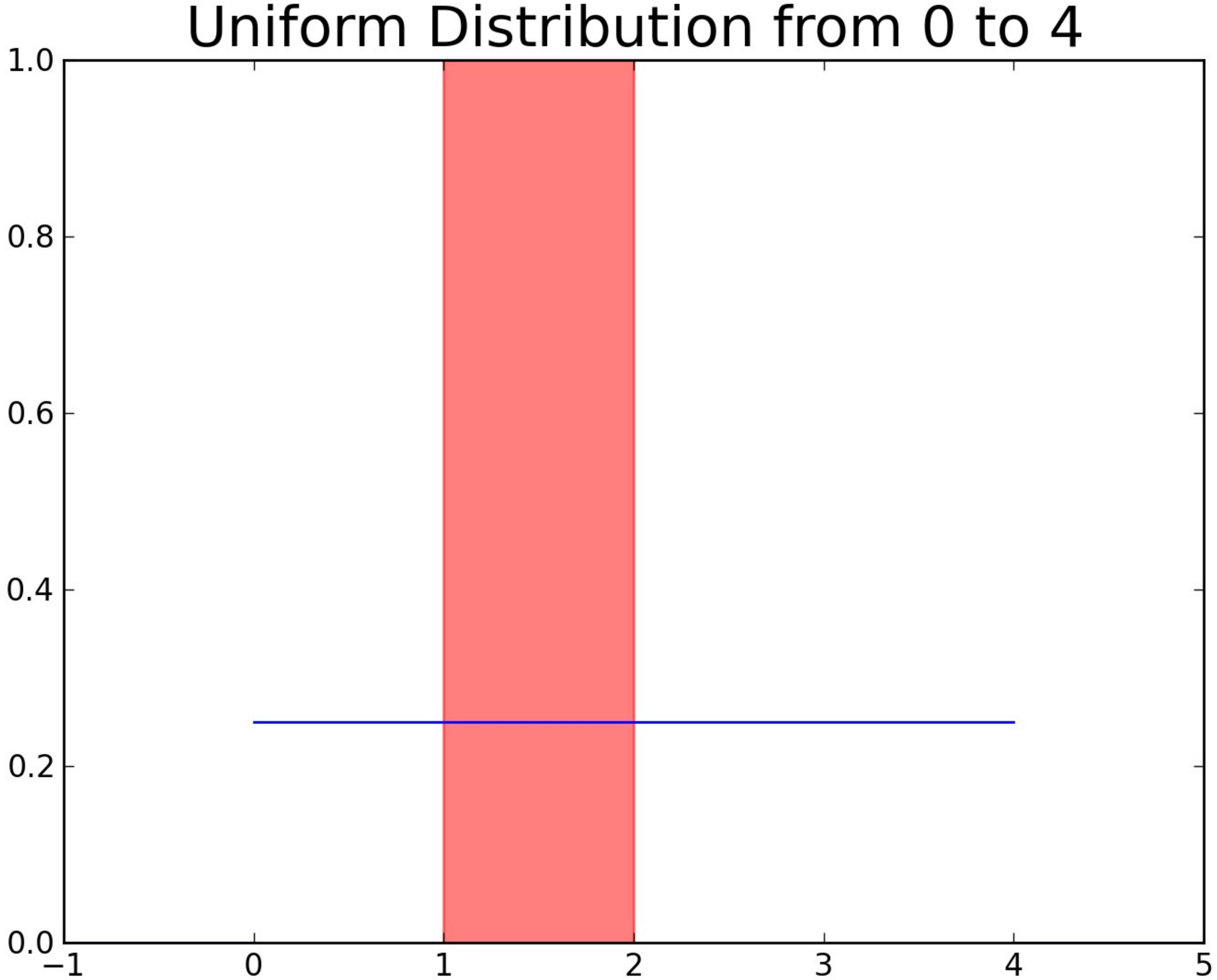
Let's be more math-y



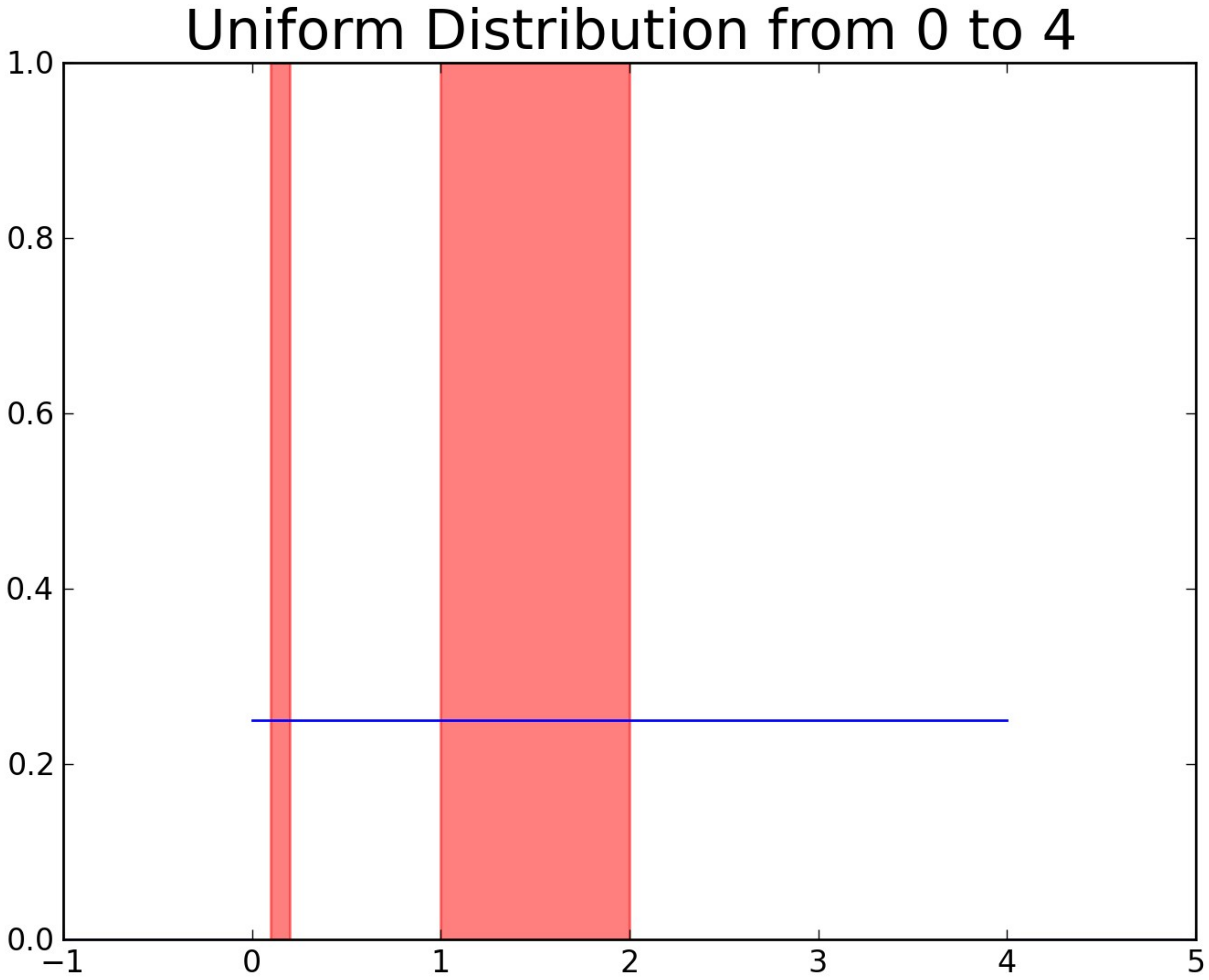
Let's be more math-y



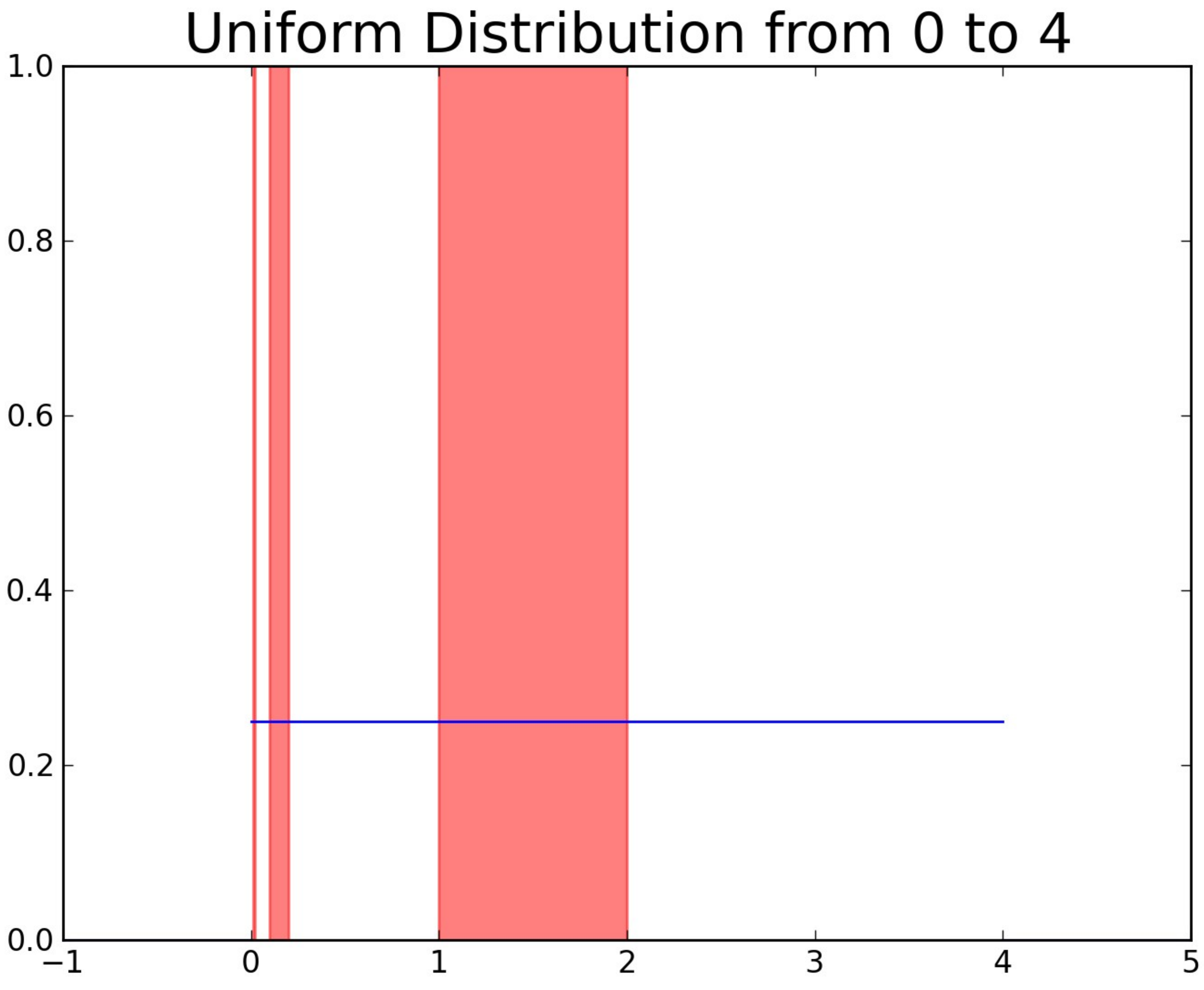
What's the probability of getting a 1 as first non-zero digit?


Uniform Distribution from 0 to 4
1.0
0.8
0.6
0.4
0.2
0.0
−1
0
1
2
3
4
5

$$P\left(1 \text{ as first non-zero digit}\right)=\int_{1}^{2} pdf\left(x\right)dx$$

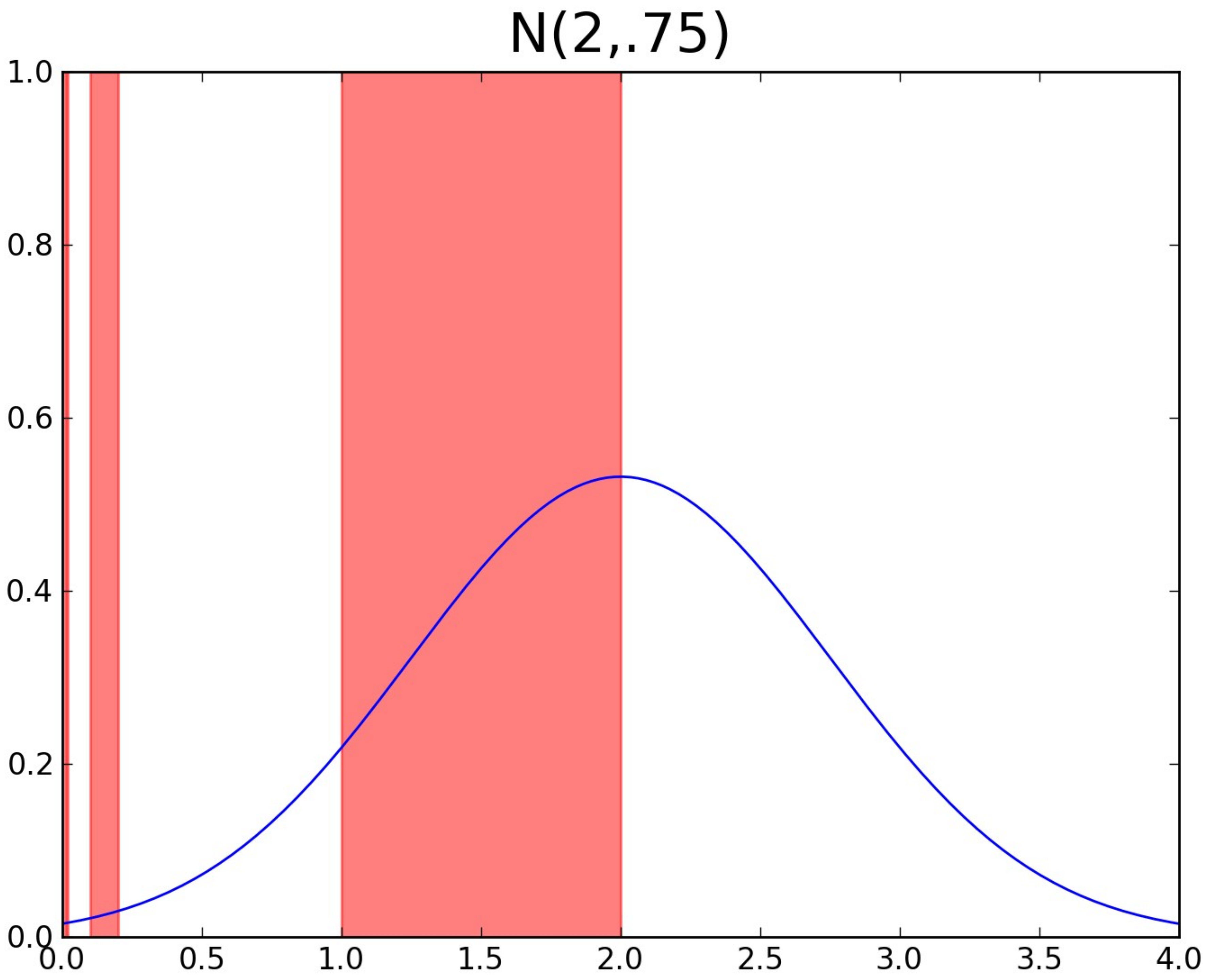

Uniform Distribution from 0 to 4
1.0
0.8
0.6
0.4
0.2
0.0
−1
0
1
2
3
4
5

$$P(1 \text{ as first non-zero digit}) = \int_{1}^{2} pdf(x)\,dx + \int_{.1}^{.2} pdf(x)\,dx$$



Uniform Distribution from 0 to 4
1.0
0.8
0.6
0.4
0.2
0.0
−1
0
1
2
3
4
5

$$P(1 \text{ as first non-zero digit}) = \int_{1}^{2} pdf(x)\,dx + \int_{.1}^{.2} pdf(x)\,dx + \int_{.01}^{.02} pdf(x)\,dx + \ldots$$



N(2,.75)
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0

- Ugh!

- ~~Ugh!~~
- Math!

Logs? Logs!

$\log_{10}$(lower bound)	$\log_{10}$(upper bound)	Δ
$\log_{10}(1) = 0$	$\log_{10}(2) = .301$	.301

Logs? Logs!

$\log_{10}$(lower bound)	$\log_{10}$(upper bound)	Δ
$\log_{10}(1) = 0$	$\log_{10}(2) = .301$	.301
$\log_{10}(.1) = -1$	$\log_{10}(.2) = -.699$	.301

Logs? Logs!

$\log_{10}$(lower bound)	$\log_{10}$(upper bound)	Δ
$\log_{10}(1) = 0$	$\log_{10}(2) = .301$	.301
$\log_{10}(.1) = -1$	$\log_{10}(.2) = -.699$	.301
$\log_{10}(.01) = -2$	$\log_{10}(.02) = -1.699$	.301
$\log_{10}(.001) = -3$	$\log_{10}(.002) = -2.699$	.301

No deep mathematical fact

Start with:

$$\log\left(\frac{A}{B}\right)=\log(A)-\log(B)$$

$$\log(1)=0$$

...

No deep mathematical fact

Start with:

$$\log\left(\frac{A}{B}\right)=\log(A)-\log(B)$$

$$\log(1)=0$$

...

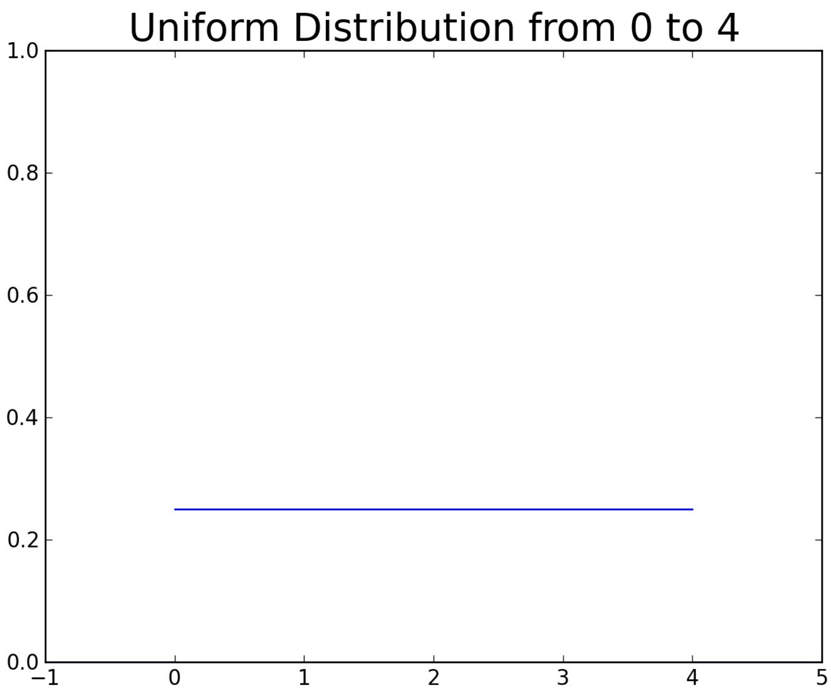
Get:

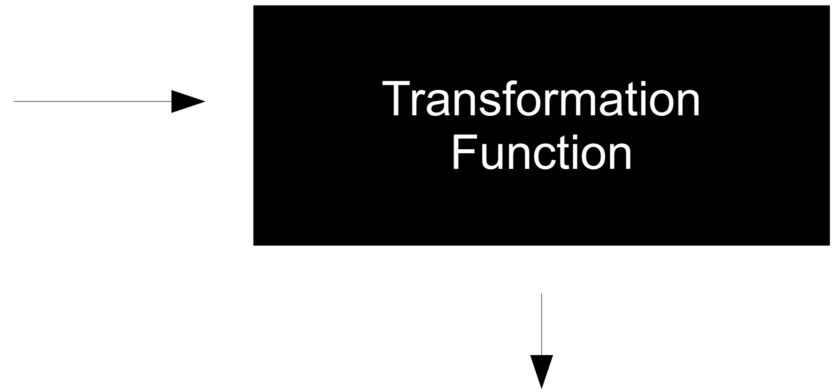
$$\log(.1) \quad = \quad \log\left(\frac{1}{10}\right) \quad = \quad \log(1)-\log(10) \quad = \quad 0-1 \quad = \quad -1$$

$$\log(.01) \quad = \quad \log\left(\frac{1}{100}\right) \quad = \quad \log(1)-\log(100) \quad = \quad 0-2 \quad = \quad -2$$

...

Convert x axis to log

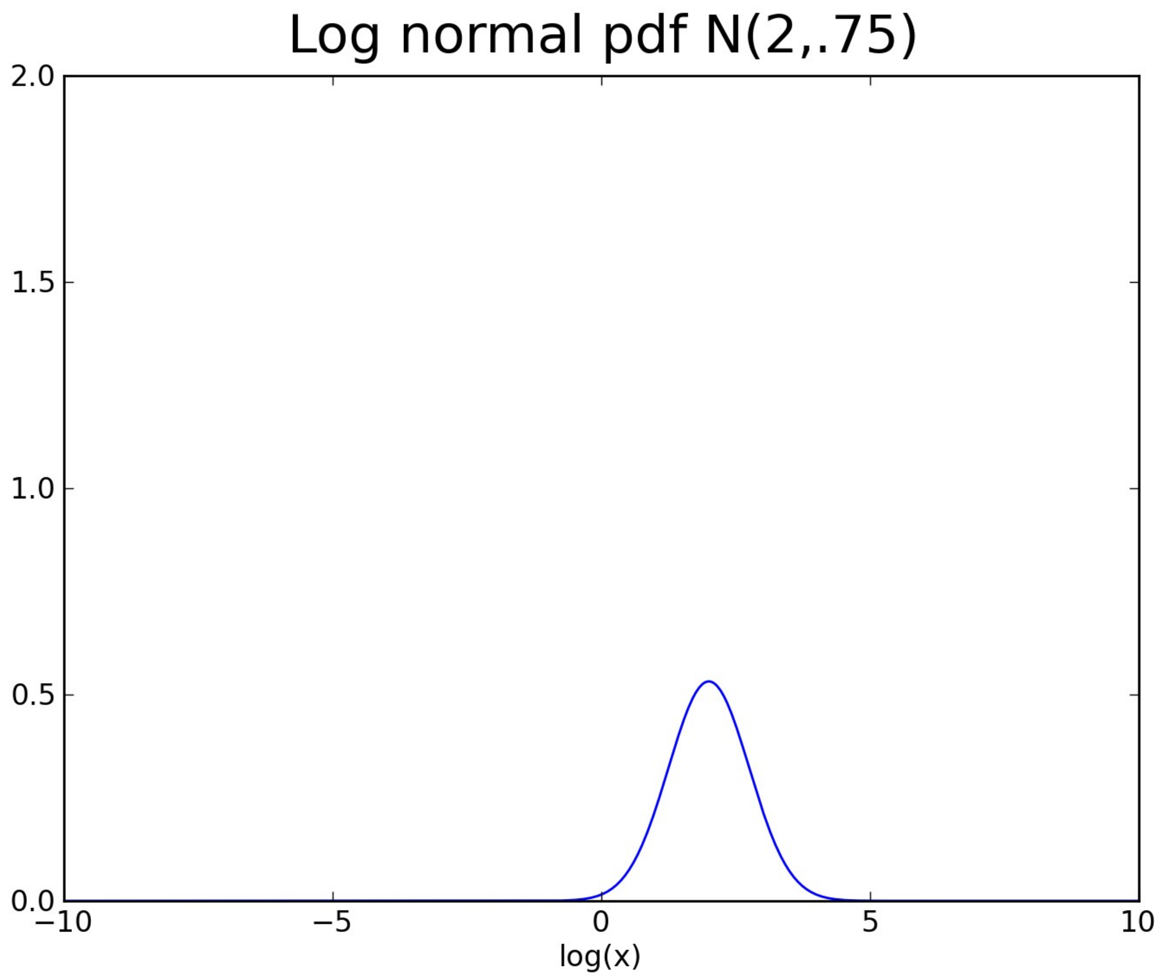




(Okay because taking log is
One-to-one
Differentiable)

Convenience

That's a little inconvenient for now. Let's just assume that we have a log-normal function. So *after* taking log, it should look like this:



Benfordian-ness of 1's (pre-log)

Before:

$$P(1 \text{ as first non-zero digit}) = \int_{1}^{2} pdf(x)\,dx +$$

$$\int_{.1}^{.2} pdf(x)\,dx +$$

$$\int_{.01}^{.02} pdf(x)\,dx +$$

...

Benfordian-ness 1's (post-log)

$P(1 \text{ as first non-zero}) =$ ... +

$P(x \in [10^{-10}, 2 \cdot 10^{-10})) +$

$P(x \in [10^{-9}, 2 \cdot 10^{-9})) +$

$P(x \in [10^{-8}, 2 \cdot 10^{-8})) +$

...

Benfordian-ness 1's (post-log)

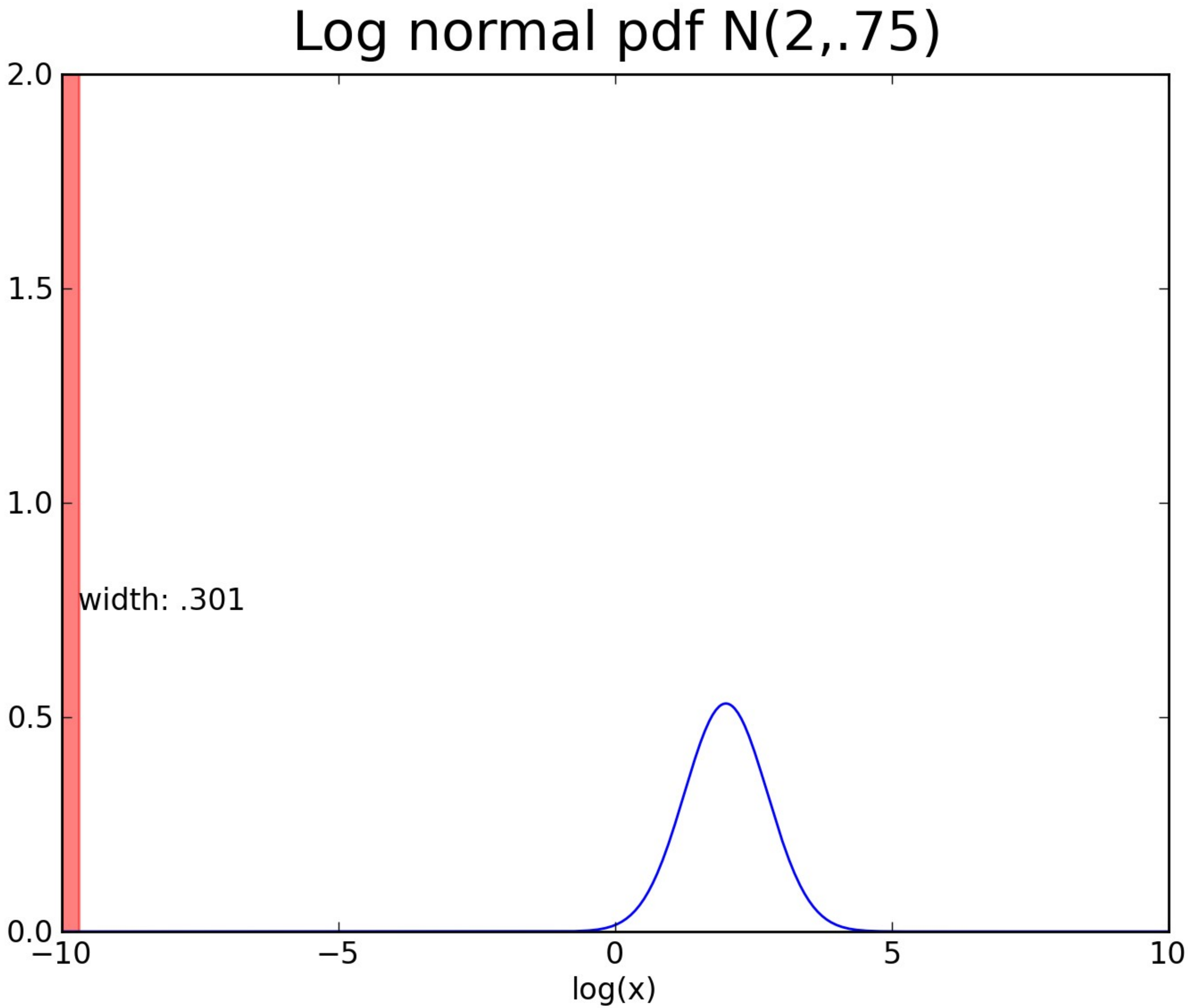
$P(1 \text{ as first non-zero}) =$ ... +

$P(x \in [10^{-10}, 2 \cdot 10^{-10})) +$

$P(x \in [10^{-9}, 2 \cdot 10^{-9})) +$

$P(x \in [10^{-8}, 2 \cdot 10^{-8})) +$

...


Log normal pdf N(2,.75)
width: .301
2.0
1.5
1.0
0.5
0.0
−10
−5
0
5
10
log(x)

Benfordian-ness 1's (post-log)

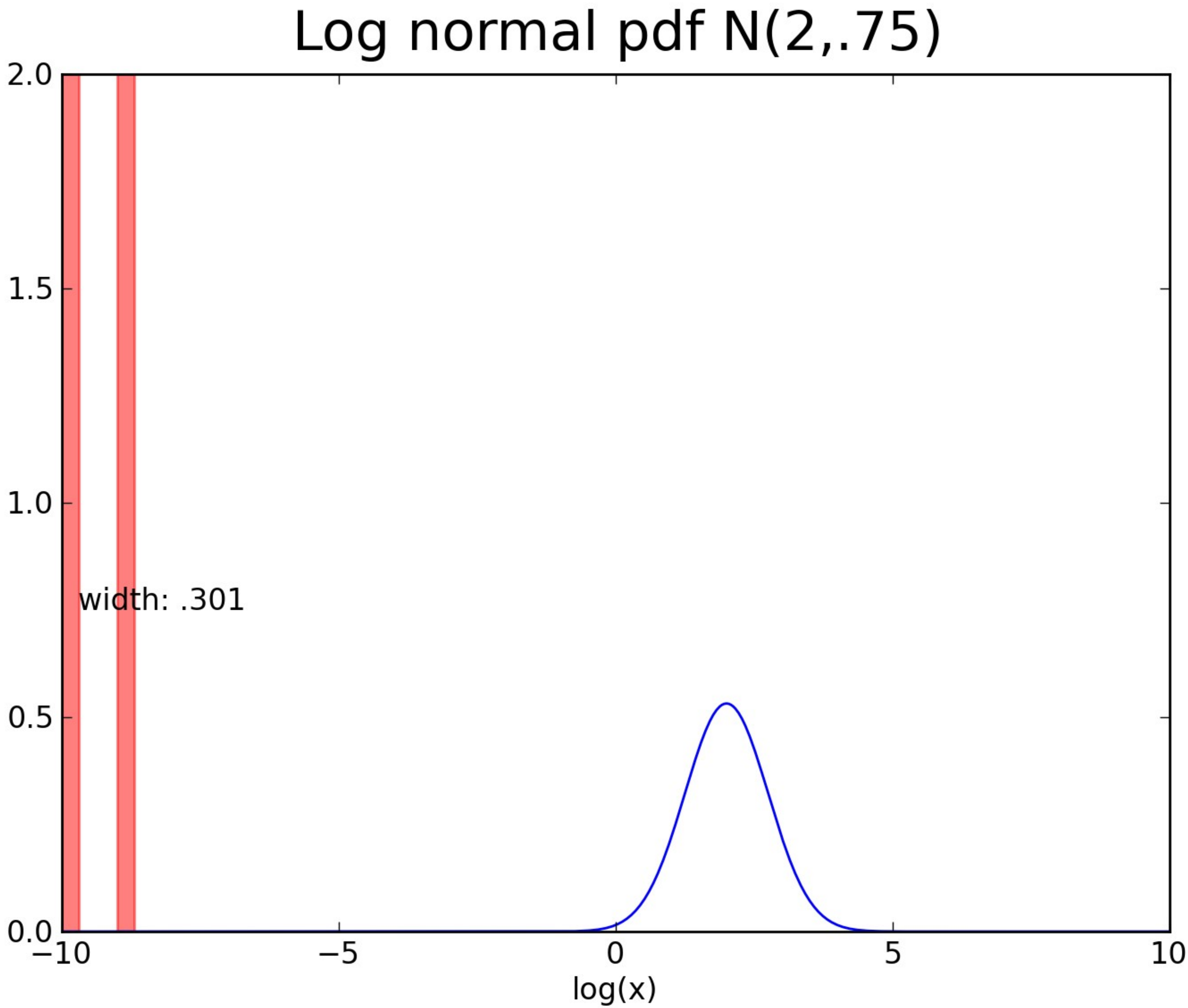
$$P(1 \text{ as first non-zero}) = \quad \ldots \quad +$$

$$P(x \in [10^{-10}, 2 \cdot 10^{-10})) +$$

$$\boxed{P(x \in [10^{-9}, 2 \cdot 10^{-9})) +}$$

$$P(x \in [10^{-8}, 2 \cdot 10^{-8})) +$$

$$\ldots$$



Log normal pdf N(2,.75)
width: .301
log(x)
2.0
1.5
1.0
0.5
0.0
−10
−5
0
5
10

Benfordian-ness 1's (post-log)

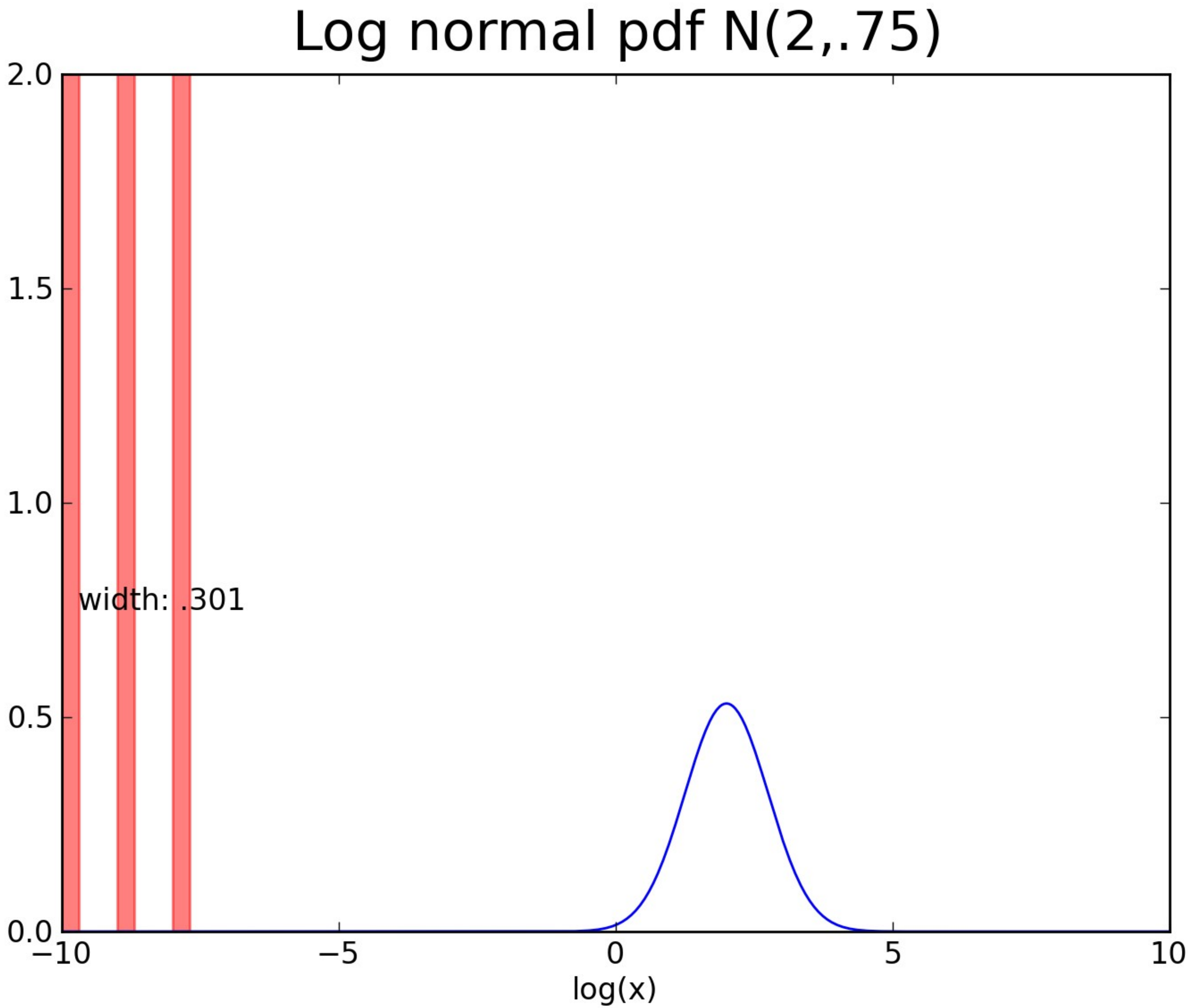
$$P(1 \text{ as first non-zero}) = \quad \ldots \quad +$$

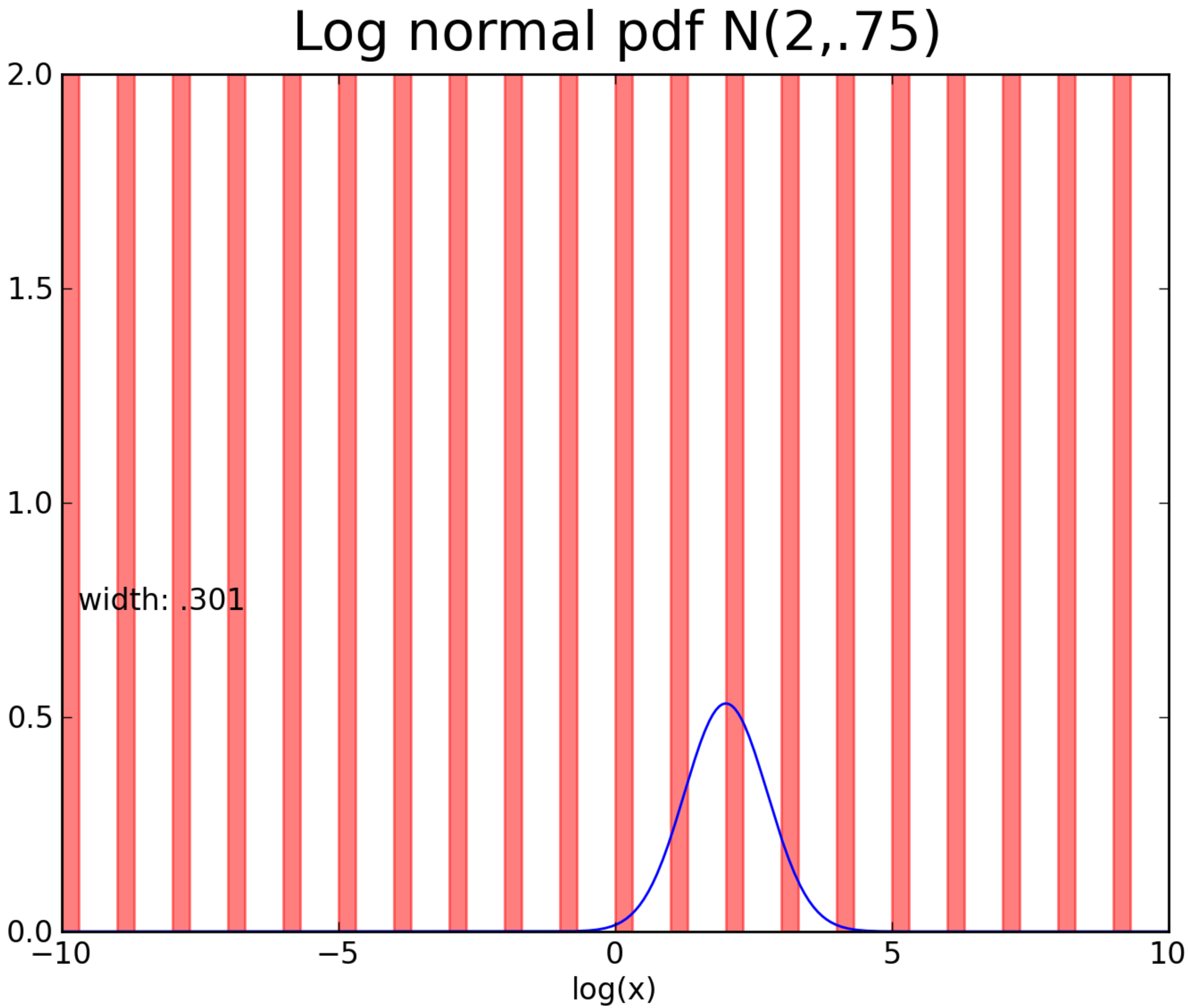
$$P(x \in [10^{-10}, 2 \cdot 10^{-10})) +$$

$$P(x \in [10^{-9}, 2 \cdot 10^{-9})) +$$

$$\boxed{P(x \in [10^{-8}, 2 \cdot 10^{-8})) +}$$

$$\ldots$$

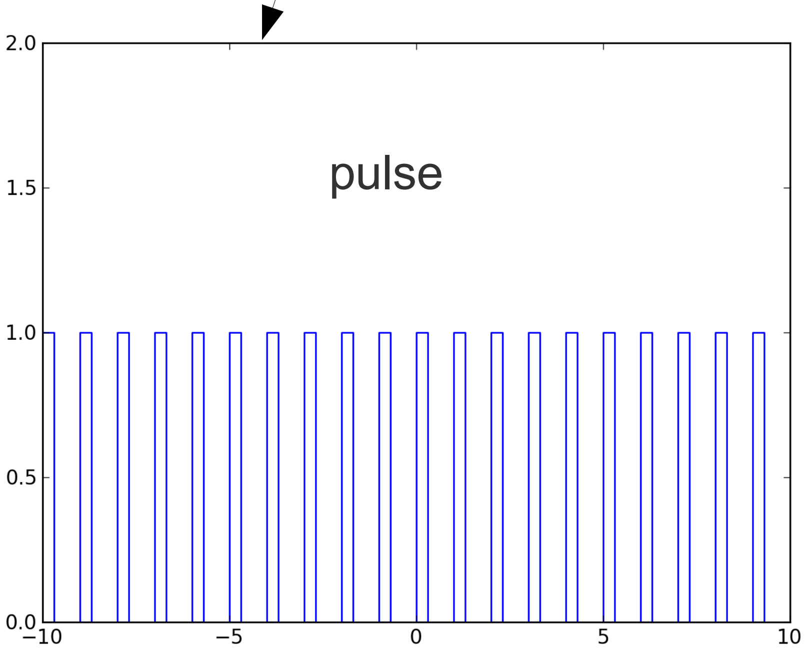

Log normal pdf N(2,.75)
2.0
1.5
1.0
0.5
0.0
width: .301
−10
−5
0
5
10
log(x)

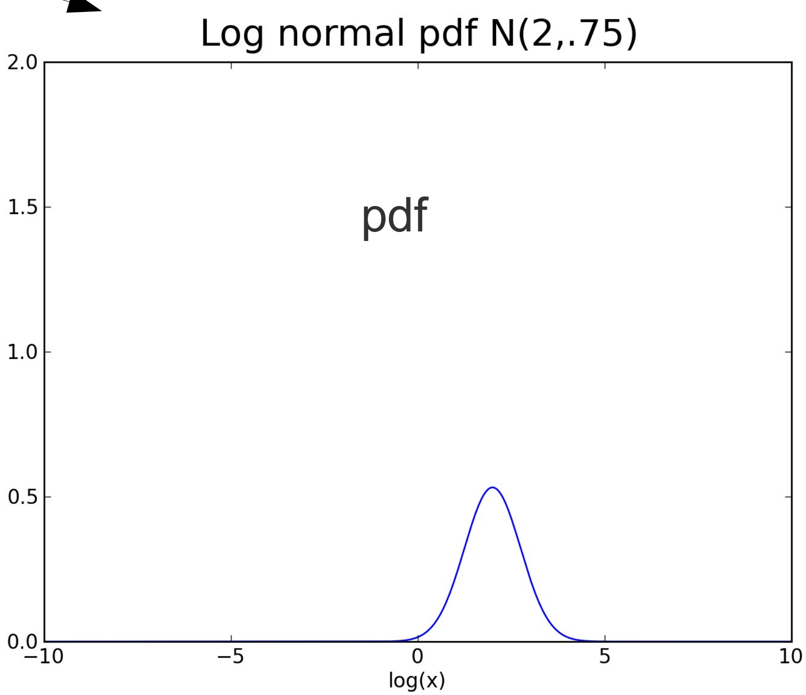


Log normal pdf N(2,.75)
2.0
1.5
1.0
0.5
0.0
width: .301
−10
−5
0
5
10
log(x)

Another way to say all that is...

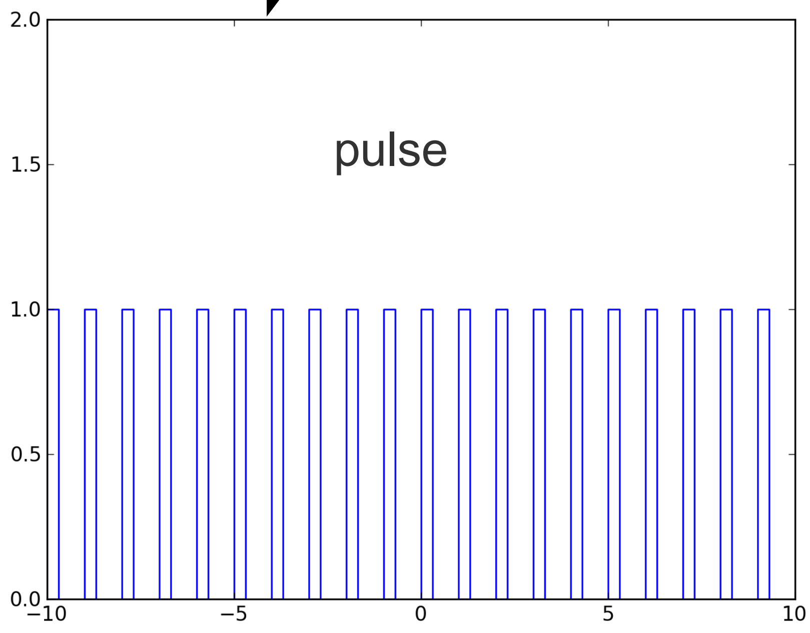
Multiply this with that and integrate result.

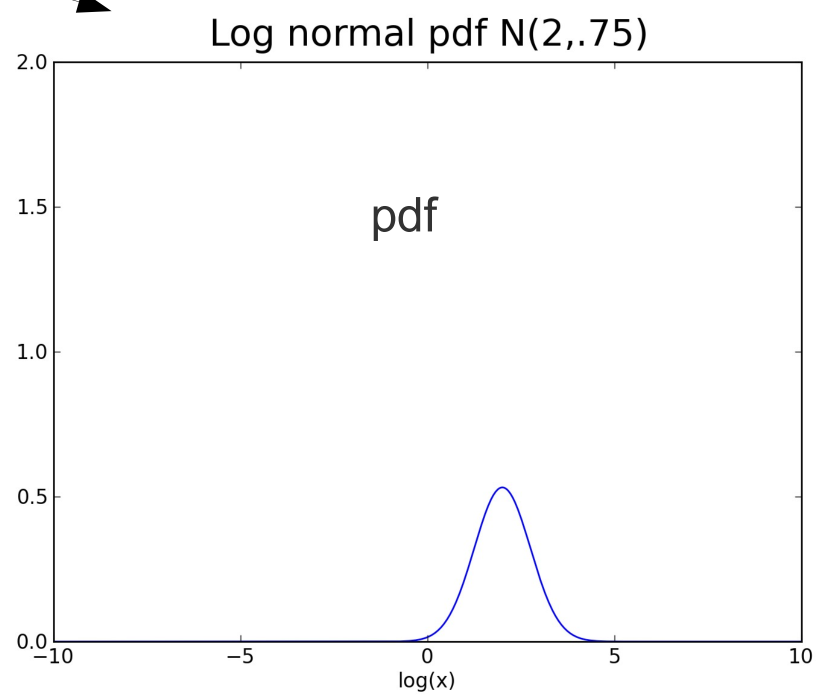




Another way to say all that is...

Multiply this with that and integrate result.





$$P(1 \text{ as first non-zero digit}) = \int_{-\infty}^{\infty} pulse(x) \cdot pdf(x)\,dx$$

What did we do?

What did we do?

```
originalData = getData(“filename.csv”)
results= [];
for s = 1:.01:1000
   testData = originalData .* s;
   results.append(fracFirstDigitOnes(testData));

plot(results)
```

What did we do?

```
originalData = getData("filename.csv")
results= [];
for s = 1:.01:1000
   testData = originalData .* s;
   results.append(fracFirstDigitOnes(testData));

plot(results)
```

$$P(1 \text{ as first non-zero digit}) = \int_{-\infty}^{\infty} pulse(x) \cdot pdf(x)\,dx$$

What do we need to do?

But that's only one part of the scaling test. We need to repeat this integral over a range of scaling constants functions.

Scaling with logs

$x \quad \text{-->} \log(x)$

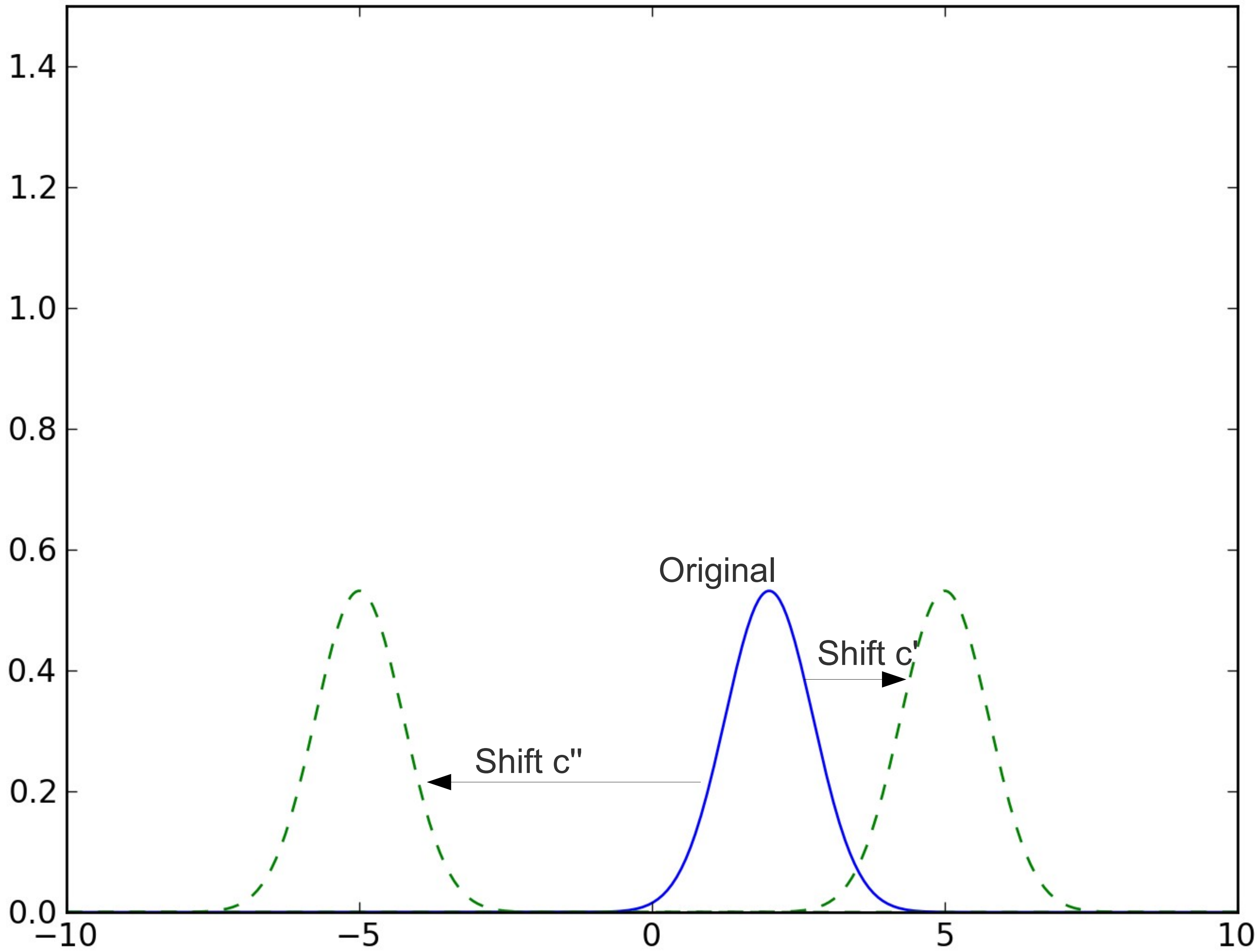
$cx \text{ -->} \log(cx) = \log(c) + \log(x)$

Scaling with logs

x   --> log(x)

cx --> log(cx) = log(c) + log(x)

Scaling by c corresponds to shifting the log distribution left or right by log(c).

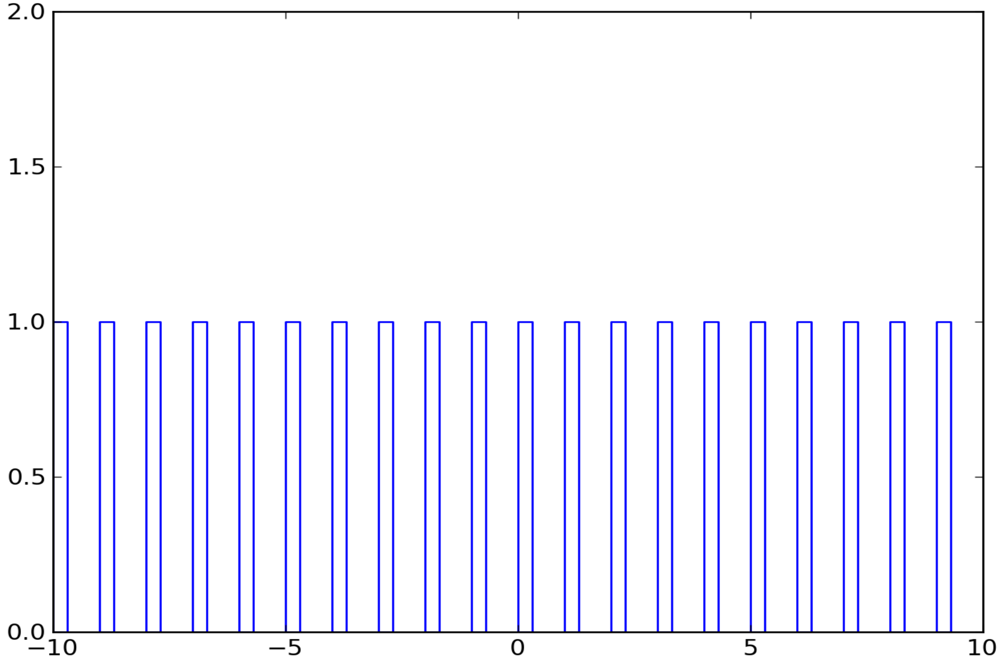

Original
Shift c'
Shift c''
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
−10
−5
0
5
10

Scaling test with math

- Sampling function remains unchanged in all of this (we never said that the sampling had to be scale invariant):

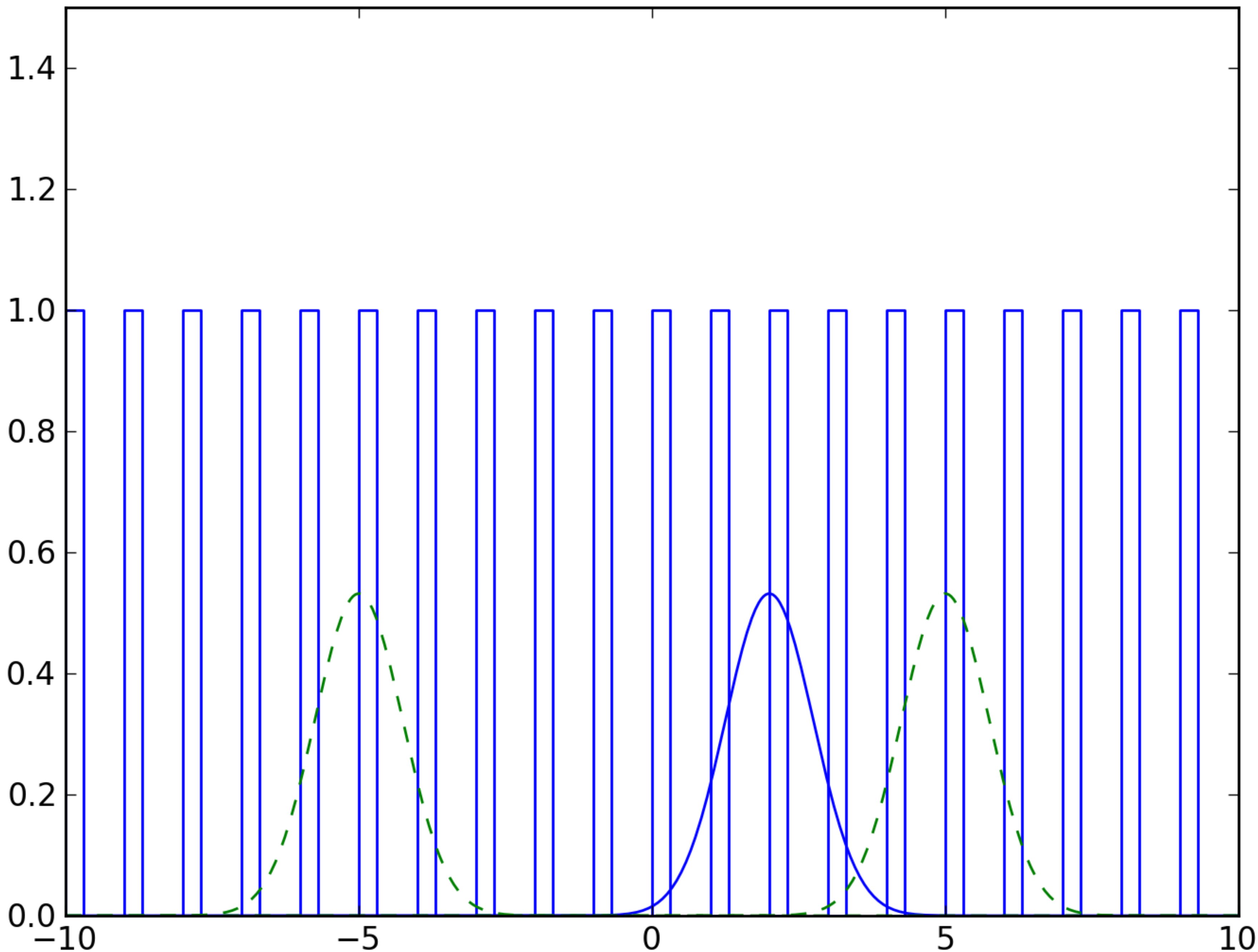
Scaling test with math

- Sampling function remains unchanged in all of this (we never said that the sampling had to be scale invariant):



Scaling test with math

- Sampling function remains unchanged in all of this (we never said that the sampling had to be scale invariant):
- Now, we're shifting pdf one way or another and integrating from $-\infty$ to ∞ depending on how much we scale.


1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
−10
−5
0
5
10

Scaling test with math

- Sampling function remains unchanged in all of this (we never said that the sampling had to be scale invariant):
- Now, we're shifting pdf one way or another and integrating from -∞ to ∞ depending on how much we scale.

$$P(1 \text{ as first non-zero digit after scaling s}) = \int_{-\infty}^{\infty} pulse(x) \cdot pdf(x - s')\,dx$$

$$where\ s' = f(s)$$

Convolution!

$$P(1 \text{ as first non-zero digit after scaling s}) = \int_{-\infty}^{\infty} pulse(x) \cdot pdf(x - s')\,dx$$

$$where\ s' = f(s)$$

$$P(1 \text{ as first non-zero digit after scaling s}) = pulse * pdf$$

Convolution!

$$P(1 \text{ as first non-zero digit after scaling s}) = \int_{-\infty}^{\infty} pulse(x) \cdot pdf(x - s')\,dx$$

$$where\ s' = f(s)$$

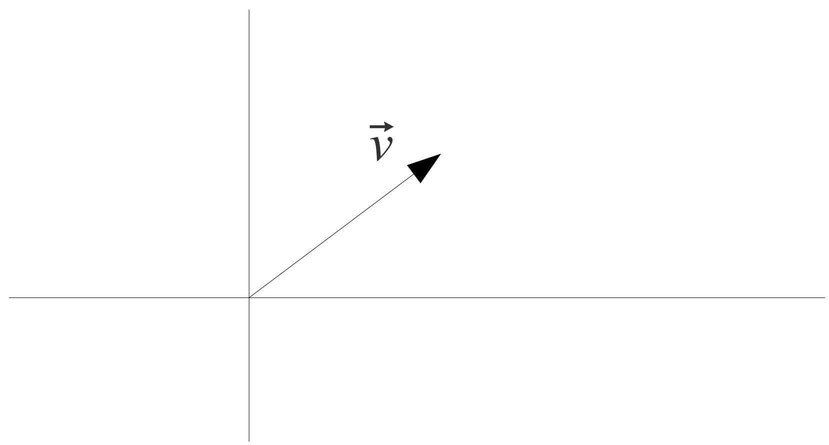
$$P(1 \text{ as first non-zero digit after scaling s}) = pulse * pdf^{-}$$

Easier to solve in frequency domain.

Got to do a *little* bit of Fourier stuff

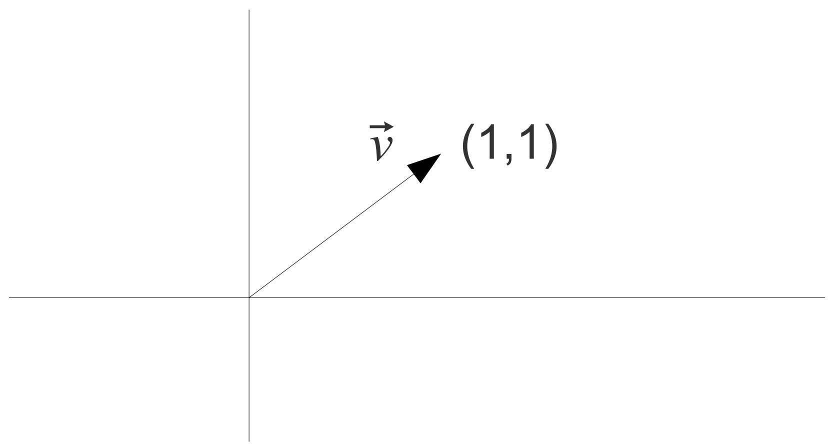
(Brief) What's the frequency domain?

If I have a vector in 2D space, that looks like this, how could you describe it?



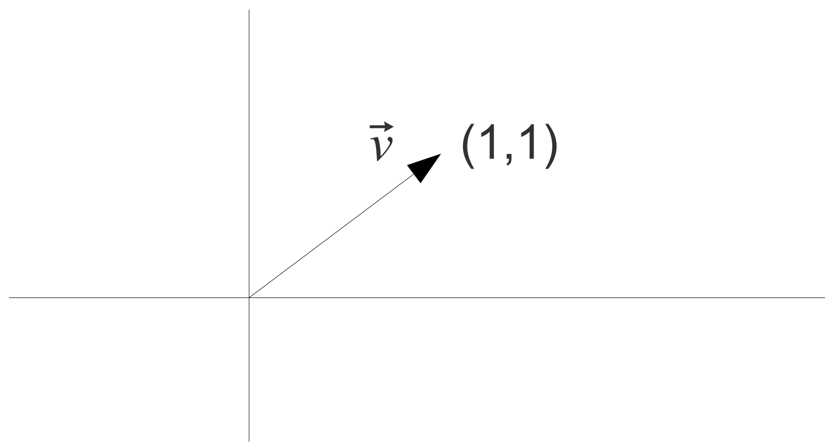
(Brief) What's the frequency domain?

If I have a vector in 2D space, that looks like this, how could you describe it?



(Brief) What's the frequency domain?

If I have a vector in 2D space, that looks like this, how could you describe it?



$$\vec{v} = 1 \cdot \hat{x} + 1 \cdot \hat{y}$$

(Brief) What's the frequency domain?

To find the frequency components of a function, you do the same thing:

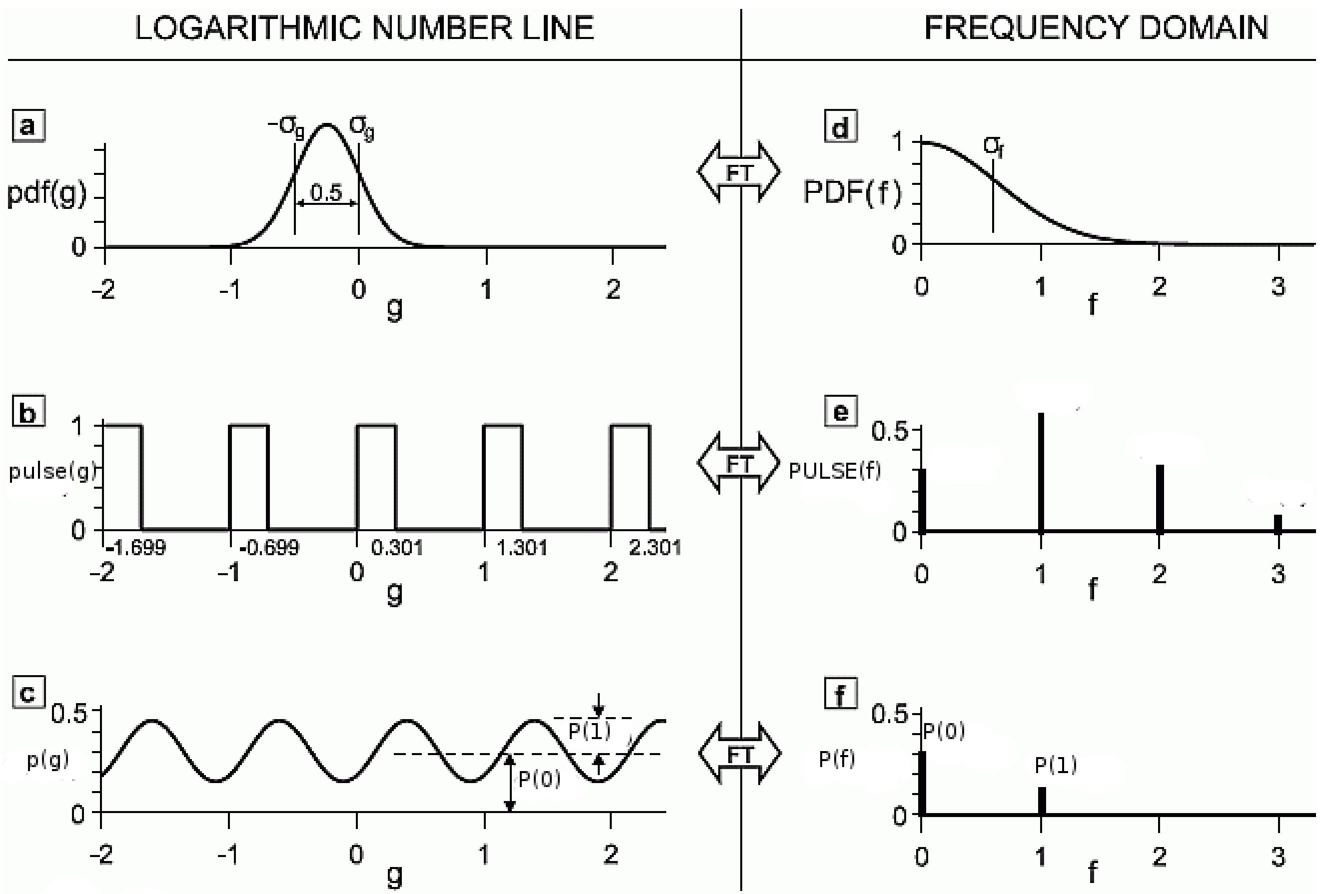
Take the inner product of the function with $\hat{f}_1$, $\hat{f}_2$, $\hat{f}_3$, etc.

All you need to know

- Convolution in time is multiplication in frequency.
- For a periodic function, p, average value of function is P(0).
- For a non-periodic function, n, integral over all values is N(0)

Back to Benford

What the multiplication in frequency actually looks like:



Minor modifications from The Scientist and Engineer's Guide to Digital Signal Processing by Steven Smith

Back to Benford

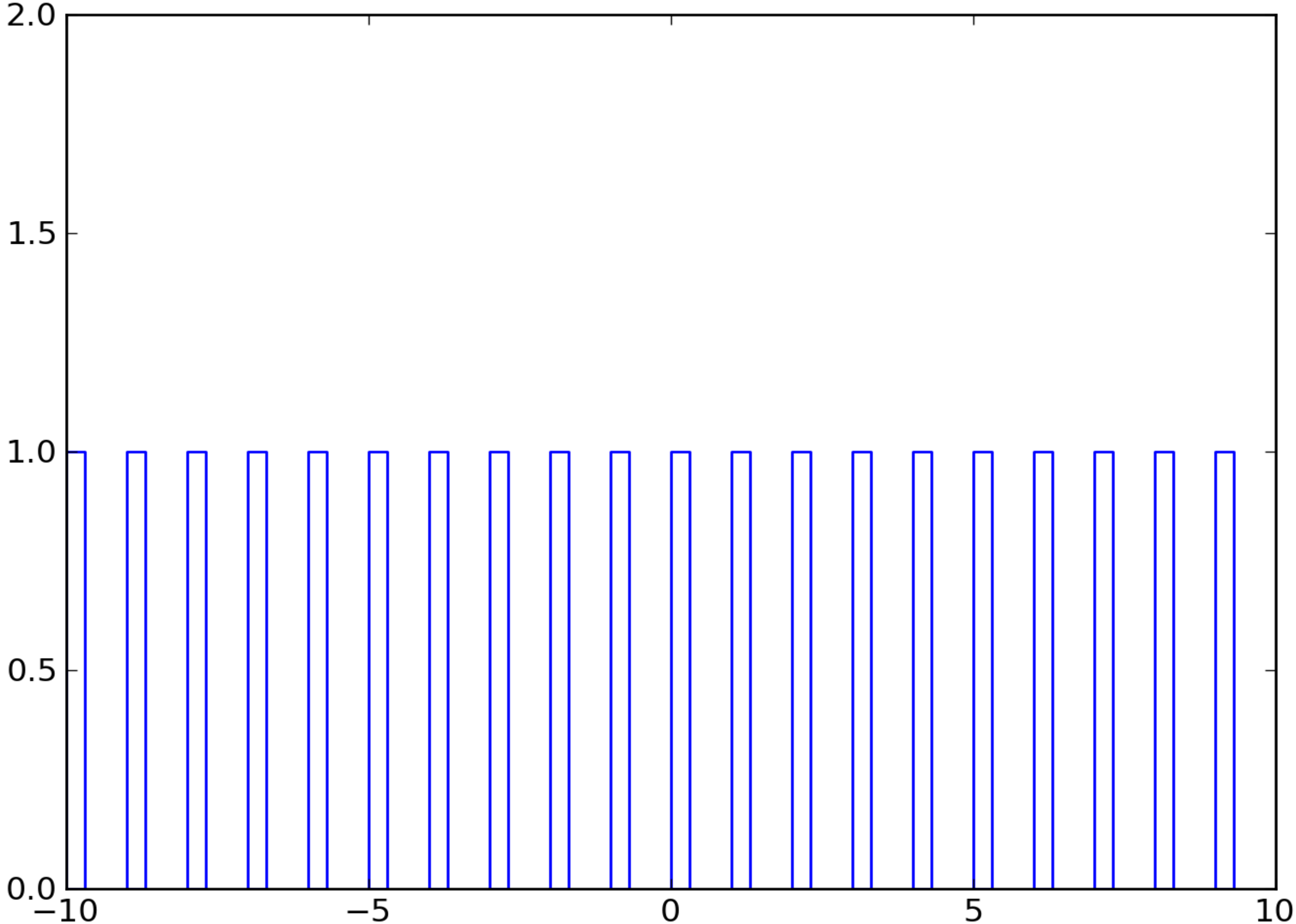
- 0 in frequency corresponds to the dc bias of our function in time.
- P(0) = PDF(0) x PULSE(0)

Back to Benford

0 in frequency corresponds to the dc bias of our function in time.

P(0) = PDF(0) x PULSE(0)

PULSE(0) = Time average over one cycle


2.0
1.5
1.0
0.5
0.0
−10
−5
0
5
10

Back to Benford

0 in frequency corresponds to the dc bias of our function in time.

P(0) = PDF(0) x PULSE(0)

PULSE(0) = .301

PDF(0) = Integral of pdf from $-\infty$ to ∞

Back to Benford

0 in frequency corresponds to the dc bias of our function in time.

P(0) = PDF(0) x PULSE(0)

PULSE(0) = .301

PDF(0) = 1

P(0) = .301

Back to Benford

0 in frequency corresponds to the dc bias of our function in time.

P(0) = PDF(0) x PULSE(0)

PULSE(0) = .301

PDF(0) = 1

P(0) = .301

If we average over all scalings for a distribution, we'd expect to see first digit 1's 30% of the time.

Slight lie from before: one more fact

- One last fact, if you stretch a function in time, you shrink it in frequency.
- If you shrink a function in time, you stretch it in frequency. (More's going on in a shorter duration, implies has to be higher frequency.)

Benford in general

- So, in general, if the distribution that we start with is “very spread out” initially, it's going to be more likely to show first-digit scale-invariance.
- *Spread out* (because we took the log) means that it should range over several orders of magnitude. Lots of data that we see does range over orders of magnitude.

Couple of other ways to think about it

- Taking lots of anti-logarithms
- Nature counts by e's
- Growth processes abound
- Show that the only distributions that behave this law need to be logarithmic.