# Statistics and Estimation

# Did you know?

The **likelihood** is the probability of the data as a function of the parameters.

# Lecture goals

Conceptual:

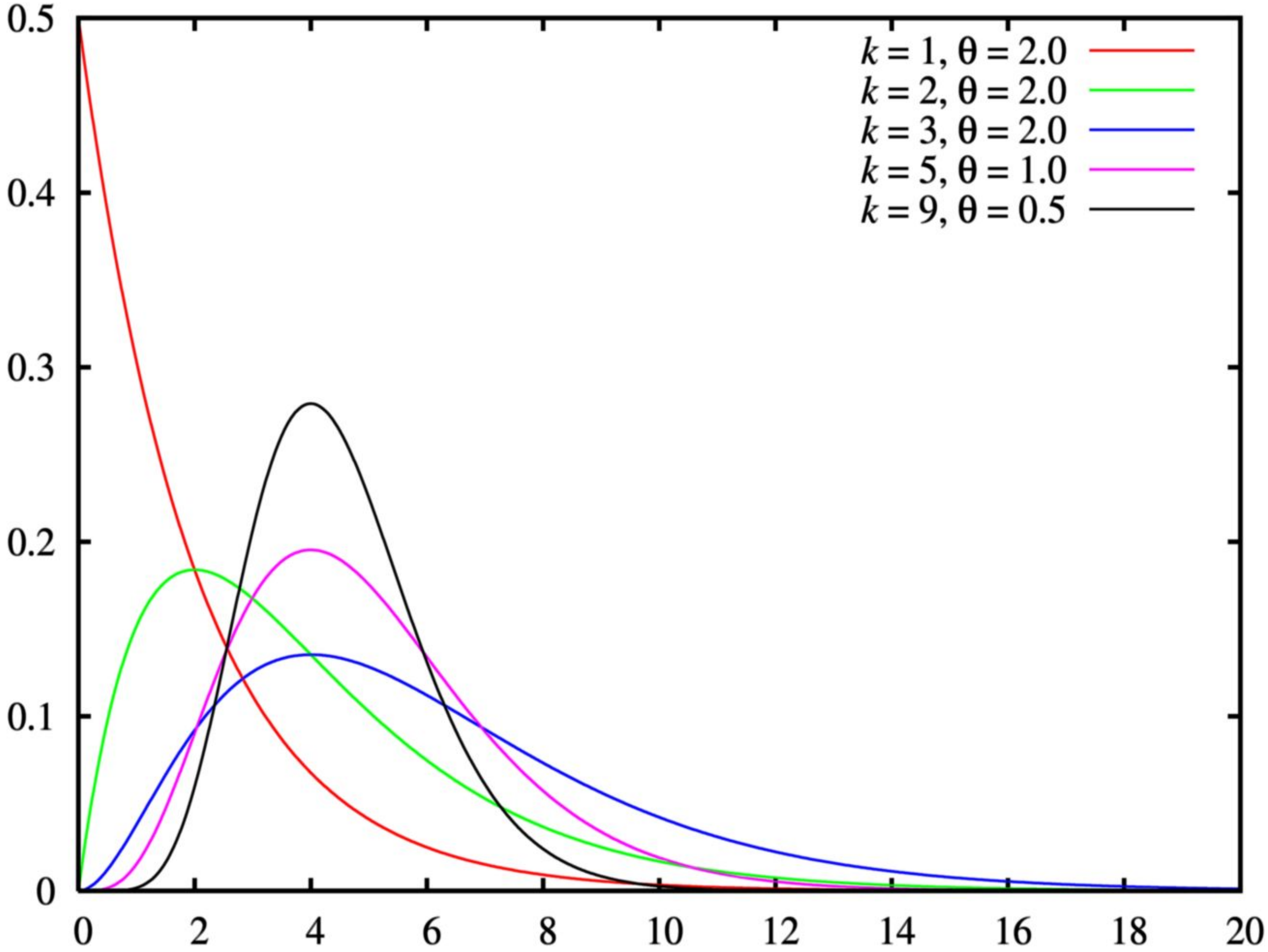
- Describe the relationship between statistics and probability
- Define a statistical model
    - Contrast it with a random variable, and a distribution
- Define closed-form vs. numerical techniques
- Define, evaluate, and optimize a likelihood function

Practical:

- Fit a model
- Plot, sample the empirical distribution of a data set
- Diagnose quality of fit visually

# Probability

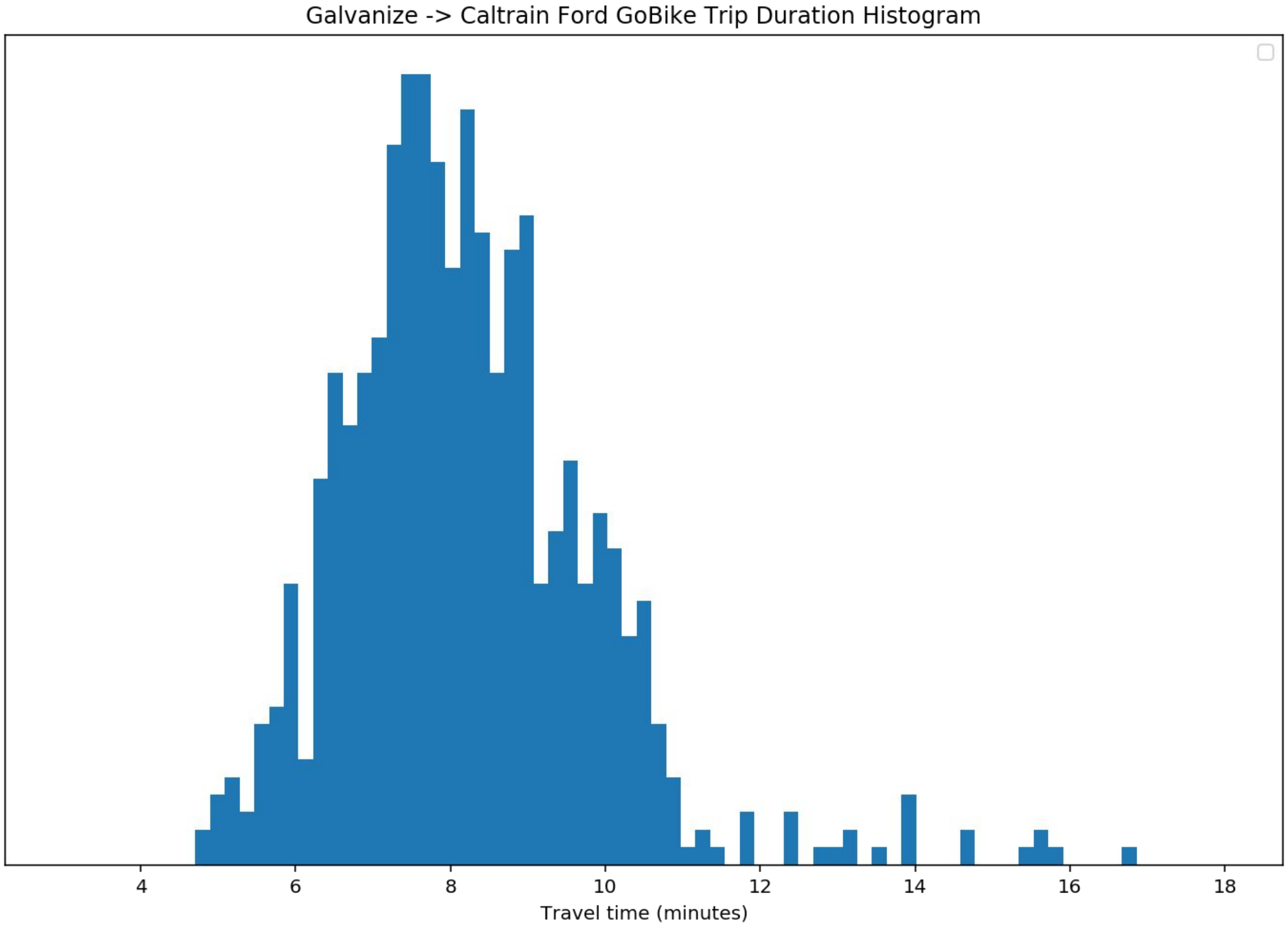
- Reason about distributions.
  - Parameters are known
- Distributions can generate data.



0.5
0.4
0.3
0.2
0.1
0
0 2 4 6 8 10 12 14 16 18 20
k = 1, θ = 2.0
k = 2, θ = 2.0
k = 3, θ = 2.0
k = 5, θ = 1.0
k = 9, θ = 0.5

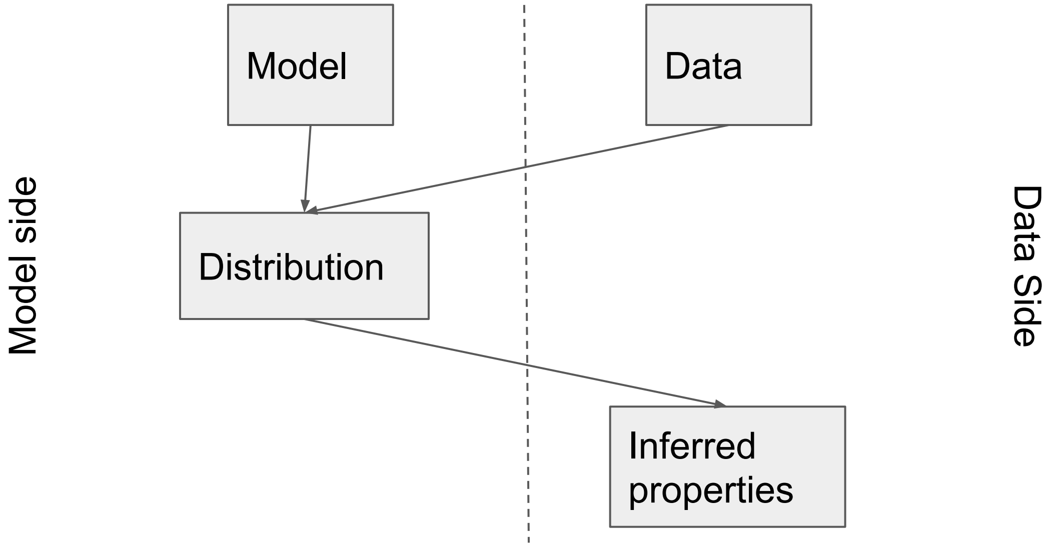
# Statistics

- Reason about data.
- Data is known.
- Data can be fit to distributions.


Galvanize -> Caltrain Ford GoBike Trip Duration Histogram
4
6
8
10
12
14
16
18
Travel time (minutes)

# How to do statistics

- Hypothesize a model
- Collect some data
- Fit the data to a model, to produce a distribution
- Use the distribution to infer properties of unseen data



# Very simple example

We have a coin.

**Model**: Each flip is a Bernoulli trial with P(heads)=p.

**Data**: Flip the coin 100 times. 70 times it comes up heads.

**Distribution**: A Bernoulli distribution with p=0.7.

**Inference**: Someone offers you a wager with even odds on a flip of said coin. You accept, bet a quantity you're willing to lose, and call heads. (Repeat as many times as you feel is ethical.)

(side note: look up “money pump”)

# A Model

- A **collection** of distributions
  - Typically with the same parameterizations
  - Examples:
    - Model: Bernoulli distribution with parameter p.
      - One parameter
    - Model: Gaussian mixture model with n components, each with a weight, mean, and variance.
      - Unbounded number of parameters
- Fitting a model
  - Means **selecting** a single distribution from the collection.
  - A fit model is not “the model”. It is a distribution with set parameters.

**Question**: What’s the difference between a random variable and a distribution with set parameters?

# Fitting a model

Means selecting a single distribution

Two broad strategies:

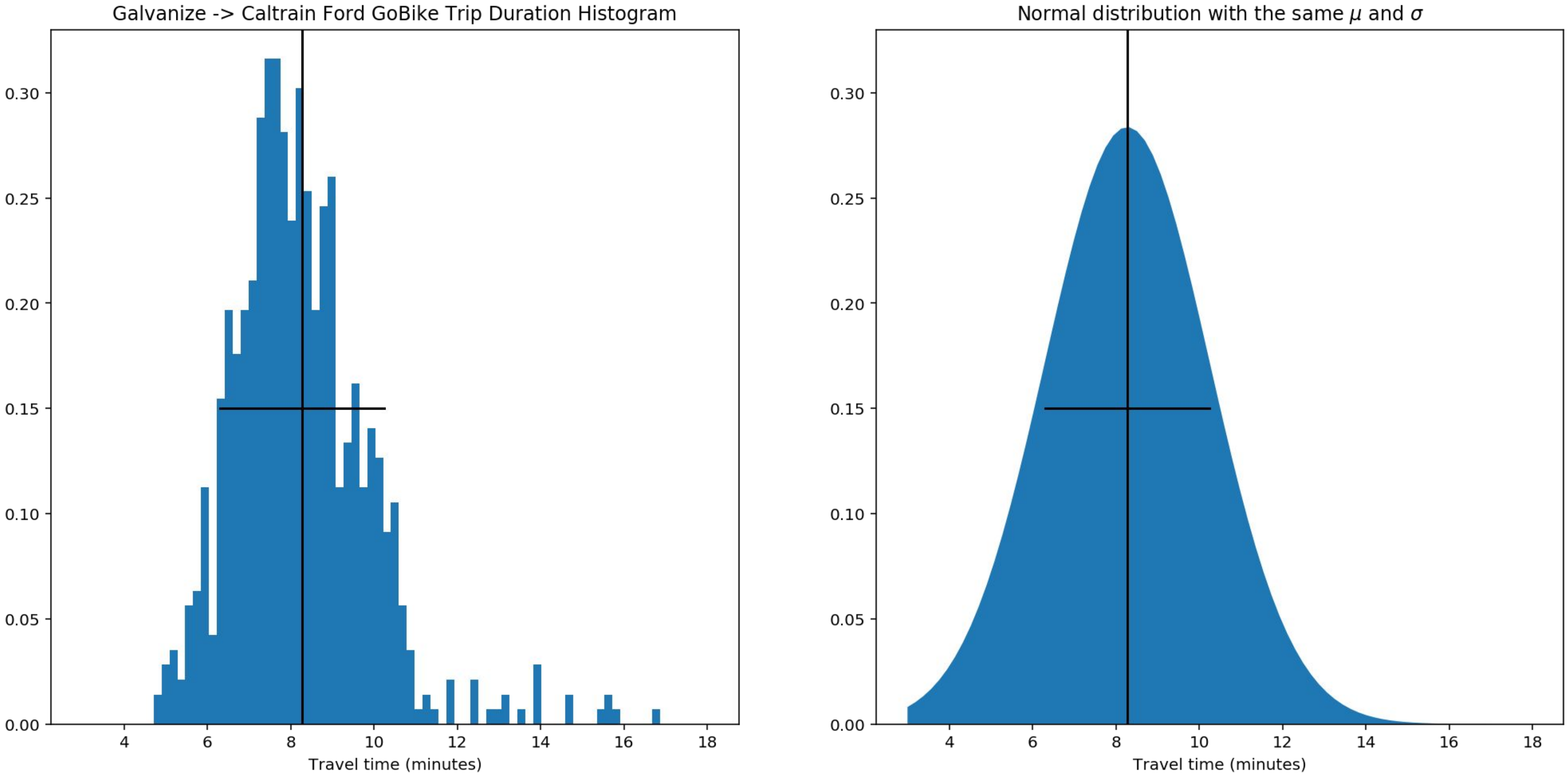
- Closed-form techniques
- Numerical techniques

# A quick aside on math vs. data science

- Closed-form techniques
  - Sometimes more principled
  - Sometimes just born out of a poverty of compute
- Numerical techniques
  - Math is hard and compute is cheap
  - Not necessarily less accurate
  - The opening up of numerical and monte carlo methods by cheap compute is what made Data Science a distinct practice.
    - Gradient descent
    - Bootstrap
    - Maximum likelihood estimation
    - Sampling from graphically-represented joint distributions

# Closed-form technique: method of moments

- Think of the distribution as a shape.
- Comparing the shape of the dataset to the shape of the distribution is computationally expensive.
- Find shape-summarizing statistics of the distribution (and dataset).
- *Simply compare those statistics.*



Galvanize -> Caltrain Ford GoBike Trip Duration Histogram
0.00
0.05
0.10
0.15
0.20
0.25
0.30
4
6
8
10
12
14
16
18
Travel time (minutes)
Normal distribution with the same $\mu$ and $\sigma$
0.00
0.05
0.10
0.15
0.20
0.25
0.30
4
6
8
10
12
14
16
18
Travel time (minutes)

# Moment: a shape-summarizing statistic

1st moment - mean - $E[X]$

2nd moment - variance - $E[(X-\mu)^2]$

3rd standard moment - skew - $E[((X-\mu)/\sigma)^3]$

4th standard moment - kurtosis - $E[((X-\mu)/\sigma)^4]$

nth standard moment - $E[(X-\mu)^n]$

# Applying the method of moments

Set

moment_n(data) = moment_n(distribution)

For as many n as you need to resolve the parameters of the distribution.

- Left is evaluated on the data
- Right is an expression in terms of the distribution’s parameters

# Fitting a uniform distribution to the bike data

moment_1(data) = moment_1(uniform dist)

$8.27 = (a+b)/2$

moment_2(data) = moment_2(uniform dist)

$1.97 = 1/12 * (b-a)^2$

2 equations, 2 unknowns; a closed-form solution exists.

# Numerical techniques

# Did you know?

The **likelihood** is the probability of the data as a function of the parameters.

I’m getting ahead of myself again

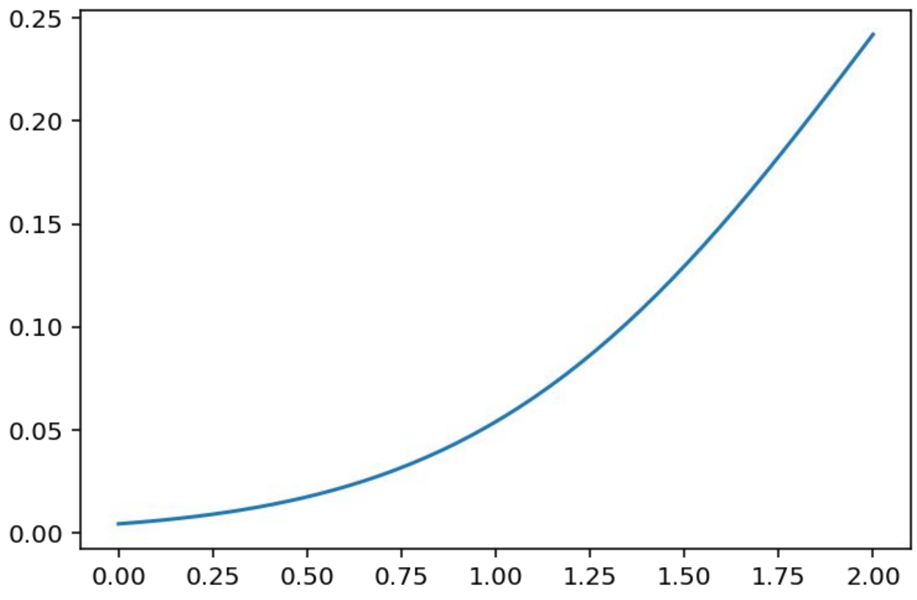
# An example

We have some data: $x=1$

And we have a distribution $N(\mu=3, \sigma=1)$

We might ask: What is $pdf(x=1 ; \mu=3, \sigma=1)$? (About 0.054).

Next question: what will happen to the probability density as we vary $x$ around 1?

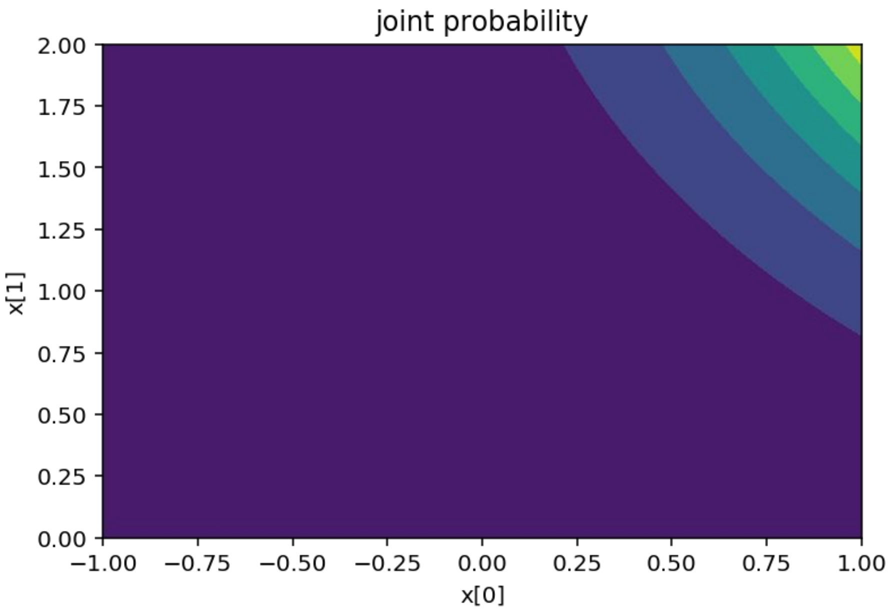



Easy enough. What's the chance of drawing a pair of values [0, 1]?

pdf(x=[0,1] ; μ=3, σ=1) = pdf(x=0 ; μ=3, σ=1) * pdf(x=1 ; μ=3, σ=1)

(about 0.00023)

What about for values around [0,1]?

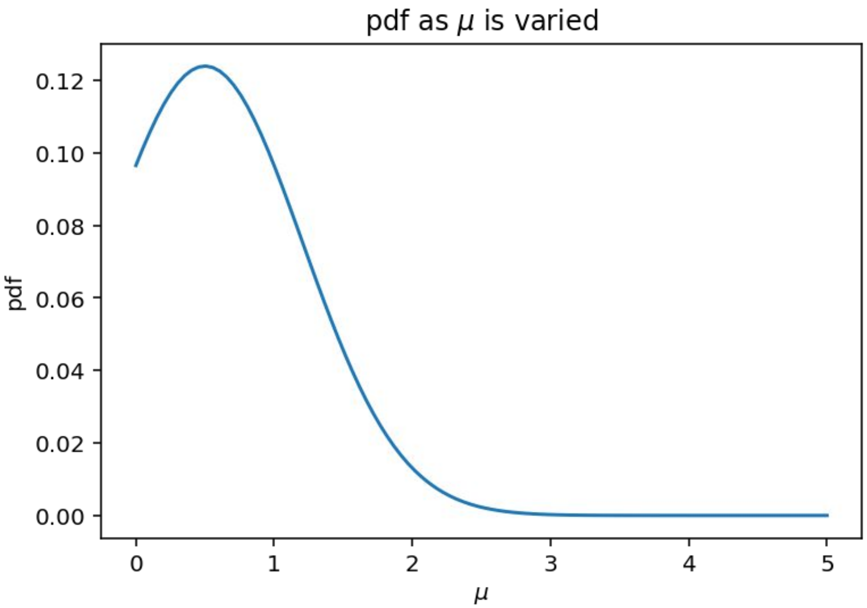



You get it. It’s a probability distribution. We can evaluate the probability of an outcome or a joint probability of a whole dataset.

# Varying the parameters instead of the data

We have our function pdf(x=[0,1] ; μ=3, σ=1)

What if we vary μ instead of the data?



# The Likelihood Function

When the data is held fixed and the P (or pdf) is evaluated as a function of the parameters, P is called the

**Likelihood function of parameters θ.**

Whereas the probability is expressed as

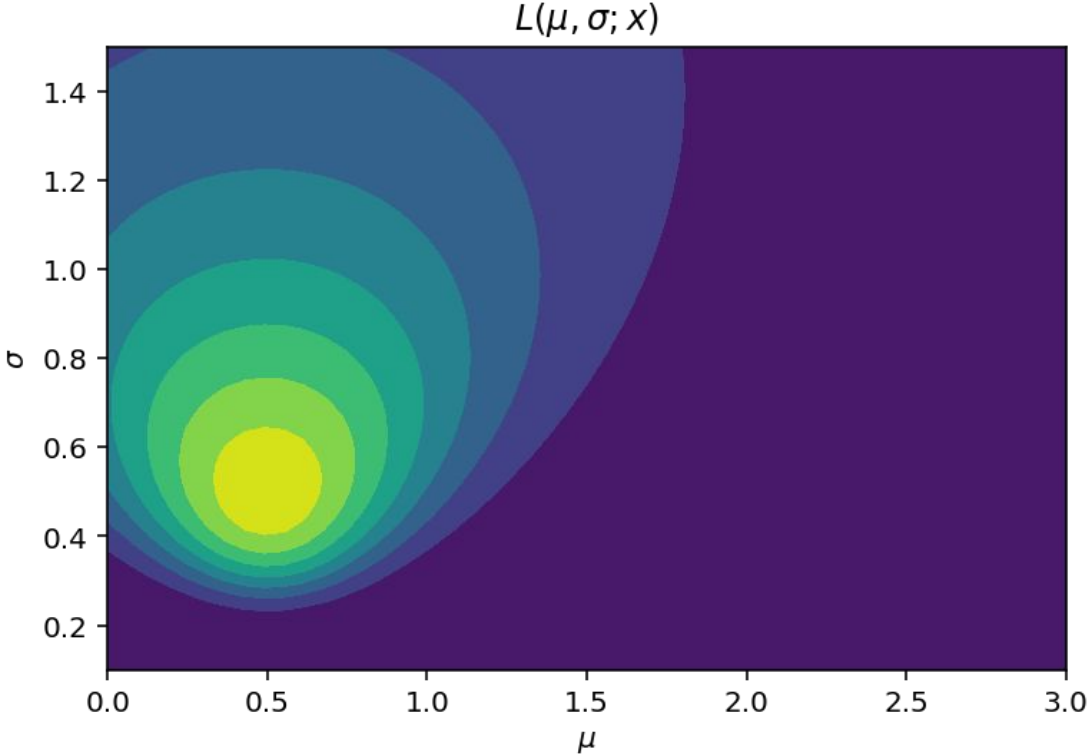
$P( x \mid \theta )$

The likelihood is sometimes expressed

$L( \theta ; x )$

# Maximizing Likelihood

In our toy example, we varied μ. Furthermore, we found a local maximum likelihood as a function of μ. We could search the joint space of (μ, σ).



# Maximum Likelihood Estimation

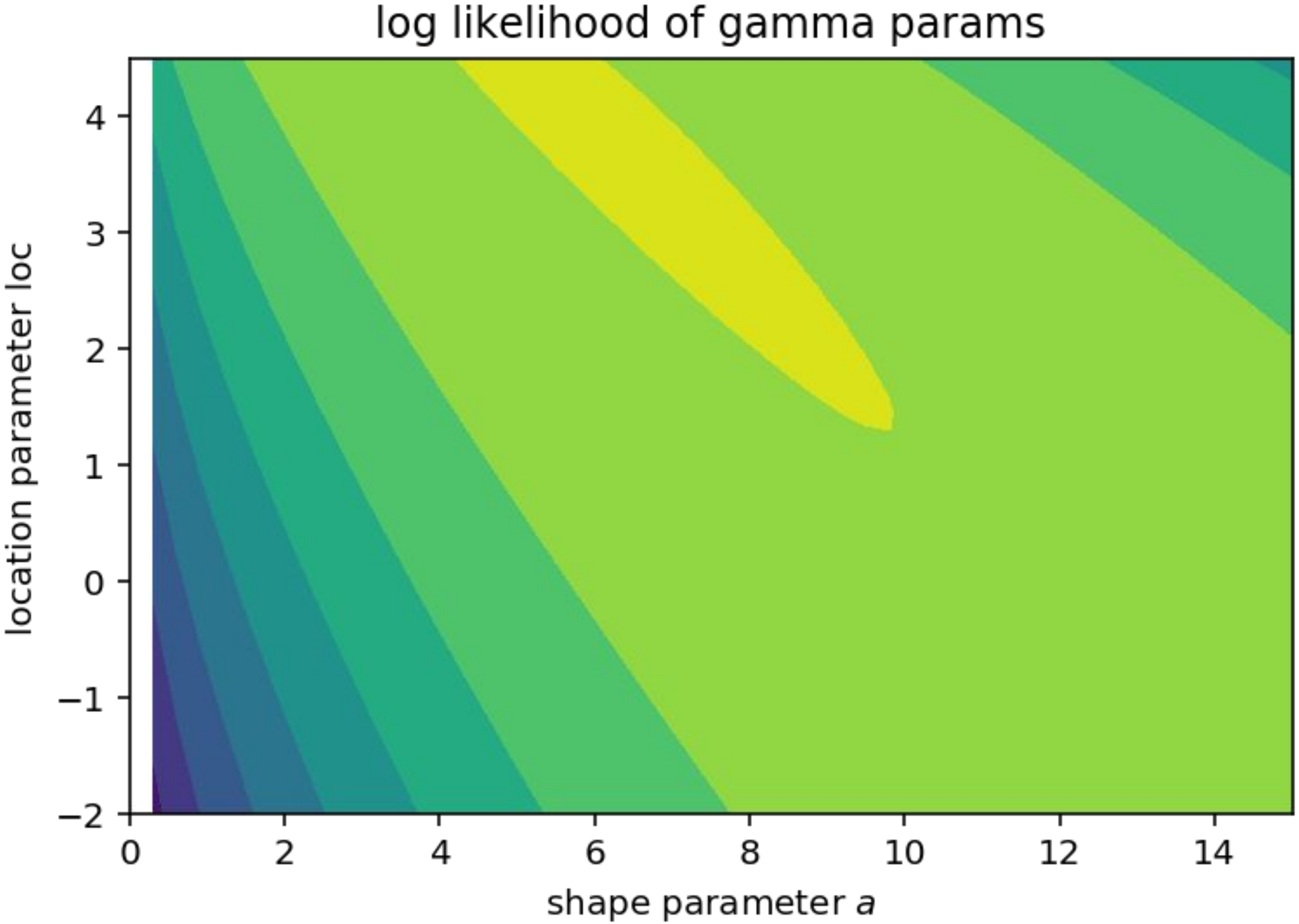
In this case, the maximum likelihood for the data x=[0,1] occurred at (μ=0.5, σ=0.5).

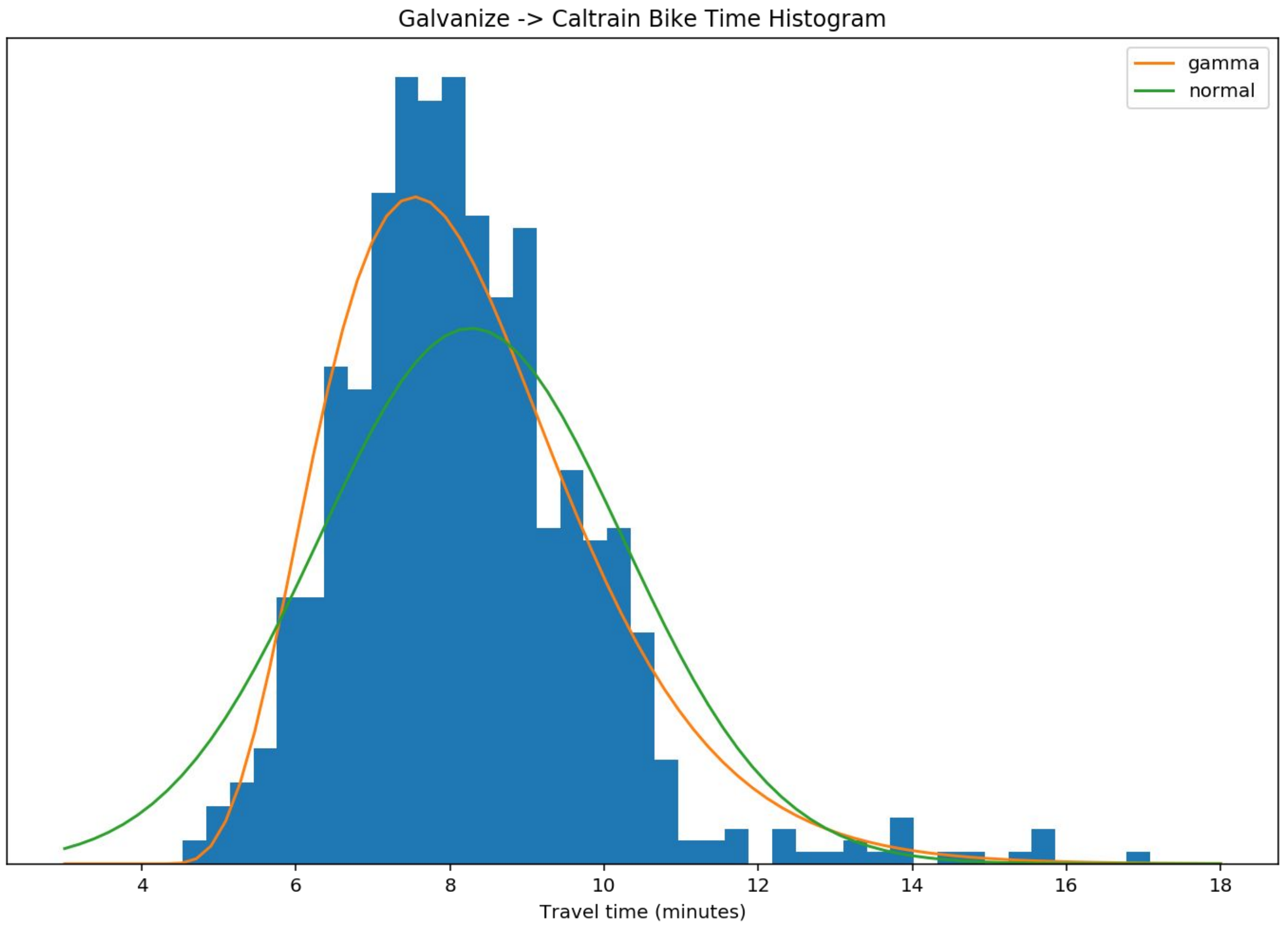
The parameter values that maximize the likelihood of the data are the **maximum likelihood estimate** for those parameters.

This provides us a **numerical technique counterpart** to the method of moments.

To apply it, we don’t need to analyze potentially very complicated distributions.

All we need is a likelihood function, and a means to maximize it.


log likelihood of gamma params
location parameter loc
4
3
2
1
0
−1
−2
0
2
4
6
8
10
12
14
shape parameter a


Galvanize -> Caltrain Bike Time Histogram
gamma
normal
4
6
8
10
12
14
16
18
Travel time (minutes)