Maximum Likelihood Estimation

by Kayla Thomas

Lecture Objectives

- Be able to explain what a statistical model is and why we use it.
- Understand that Maximum Likelihood Estimation (MLE) is a method used in statistics to determine statistical model parameters from data.
- Be able to use MLE to estimate model parameters given a statistical model and data.
- Make you comfortable explaining MLE and using it.

Review

What is the difference between probability and statistics?

1. In ----blank---- we know the parameters of a distribution (associated with some random variable), and we would like to study properties of data generated from that distribution.

2. In ----blank--- we have data generated from a random variable, and we would like to infer/estimate properties of its distribution.

Review

What is the difference between probability and statistics?

1. In **probability** we know the parameters of a distribution (associated with some random variable), and we would like to study properties of data generated from that distribution.

2. In **statistics** we have data generated from a random variable, and we would like to infer/estimate properties of its distribution.

Review

What would you be using (probability or statistics) to answer the following questions?

1. A coin is flipped 100 times, and comes up heads 55 times. Is it fair?

2. You go to play craps in Vegas, using 2 fair dice. What's the chance you'll roll snake eyes (1 and 1) on your first roll?

Review

What would you be using (probability or statistics) to answer the following questions?

1. A coin is flipped 100 times, and comes up heads 55 times. Is it fair?

Statistics

2. You go to play craps in Vegas, using 2 fair dice. What's the chance you'll roll snake eyes (1 and 1) on your first roll?

Probability

Review

Would you use statistics or probability to answer this one?

While getting ready for work, after waking up at 6:00, you observe busses stopping at your bus stop at 6:12, 6:15, 6:20, 6:21, and 6:30, after which you leave to the bus stop at 6:33. What is the probability you will have to wait longer than 5 minutes for the next bus?

Review

Would you use statistics or probability to answer this one?

While getting ready for work, after waking up at 6:00, you observe busses stopping at your bus stop at 6:12, 6:15, 6:20, 6:21, and 6:30, after which you leave to the bus stop at 6:33. What is the probability you will have to wait longer than 5 minutes for the next bus?

Both!!

The Process

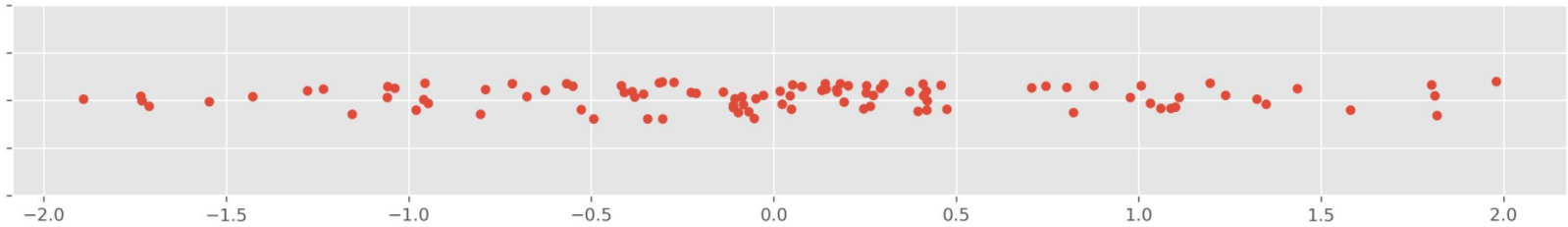
1. Use domain knowledge/creativity to create a statistical model for your data, this is a catalogue of possible ways that the data may have been generated.
2. Fit the statistical model to the data. This selects exactly one of the data generating processes from your catalogue and determines its parameters.
3. Use the fit model to compute the thing you want. Now we are using it in a probabilistic way.

Maximum Likelihood Estimation (MLE) - what most of this lecture is about - helps with step 2.

Definitions

A **statistical model** is a collection of random variables, each of which is hypothesized to possibly have generated the data. Different random variables in the collection are usually distinguished by **parameters**.

For example, we might have data that we think was generated by a normal distribution. So from the data, we would need to determine the parameters that defines the normal distribution (mean, standard deviation).



Definitions

Fitting a statistical model to data is any process that combines a model with data, and uses the data to select one and only one random variable from the model. This often takes the form of **determining the parameters for one and only one of the random variables in the model**. These estimated values of the parameters are called **parameter estimates**.

Definitions

Fitting a statistical model to data is any process that combines a model with data, and uses the data to select one and only one random variable from the model. This often takes the form of **determining the parameters for one and only one of the random variables in the model**. These estimated values of the parameters are called **parameter estimates**.

Warning from a statistician: This terminology is one of the most abused in all of statistics. Formally, the model is a **collection** of possible data generating process, but everyone also refers to the single object you get after fitting the model to data as "the model". **Seriously everyone does this**, but it's technically incorrect.

Statistical Model Example

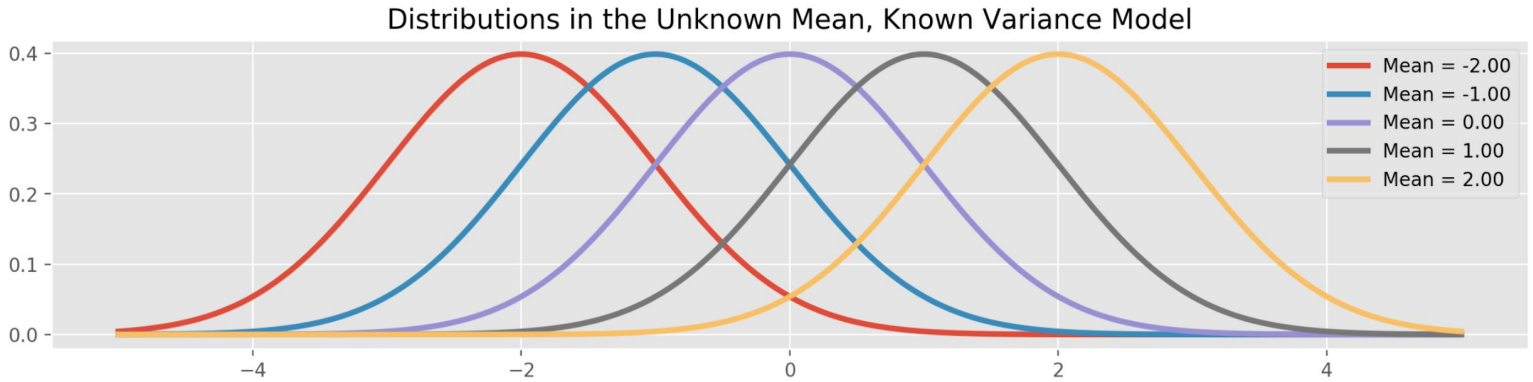
Let's fit a model to data as part of a thought experiment. We suspect that our data is generated from a normal distribution with **unknown mean**, but we know the variance is 1 (this is not something you can generally know, we're assuming it for simplicity in the thought experiment. We have only one parameter to estimate: the **mean**).

Then, our **statistical model** is the following collection of random variables:

$$M=\{\text{Normal}(\mu,1)\,|\,\mu\in\mathbb{R}\}$$

Statistical Model example

The **parameter** at play is the unknown mean of the normal distribution (since we are assuming we know the variance, that's not a parameter).



By **fitting the model to data** we mean any process that selects exactly one of these distributions. This reduces, in this case, to selecting the mean parameter μ of the unknown normal distribution.

The Eyeball Method

Optimization methods, such as Maximum Likelihood Estimation, have a mathematically rigorous way to select the right parameter.

First, let’s trying to build our intuition about the process. So we're going to first illustrate the selection of parameter using the the slightly hand-wavy eyeball method.

The Eyeball Method

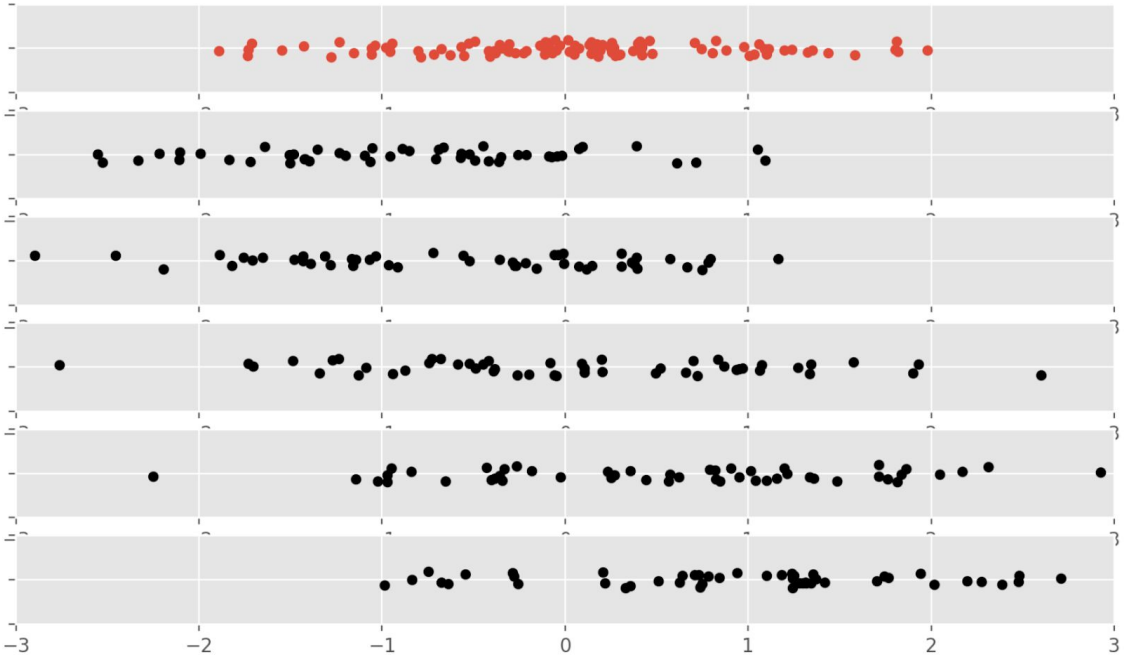
To do this we will

1. Sample new data sets - same n, from normal distributions, varying mean
2. Pick the distribution that best resembles our data

Using our eyeballs, we're going to compare data generated by normal distributions of varying means to the data that we got, and pick whichever mean gives us data most similar to the data we got.

The Eyeball Method

Which mean gives data most similar to the True Data?



True Data

Sampled From N(-1, 1)

Sampled From N(-0.5, 1)

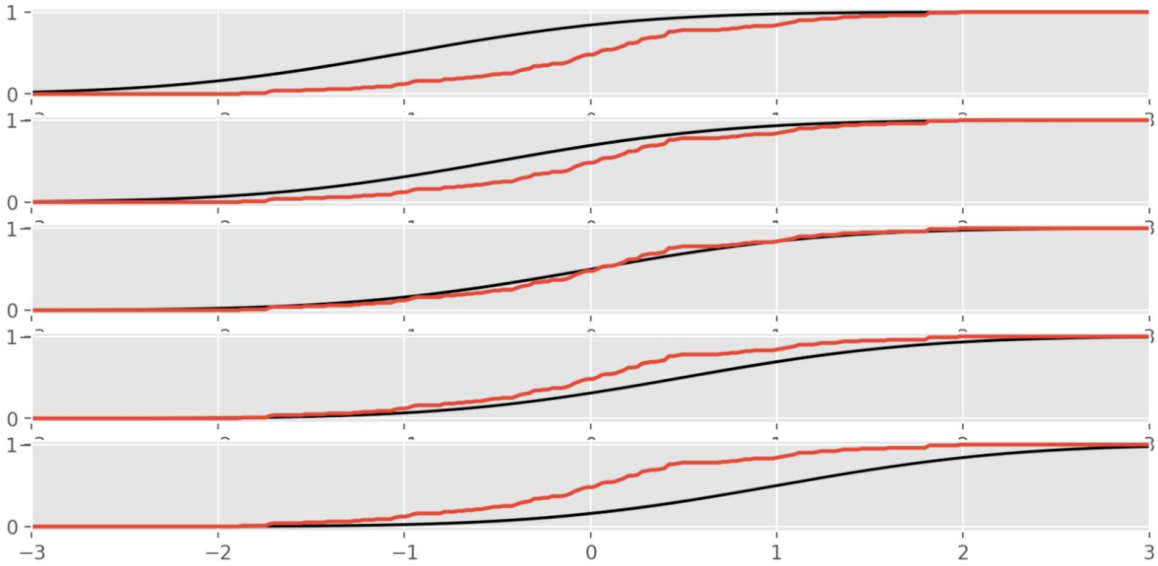
Sampled From N(0, 1)

Sampled From N(0.5, 1)

Sampled From N(1, 1)

The Eyeball Method

Another way to eyeball is to compare the empirical CDF to the actual CDFs of the candidate distributions.



CDF of Normal(-1.00, 1)

CDF of Normal(-0.50, 1)

CDF of Normal(0.00, 1)

CDF of Normal(0.50, 1)

CDF of Normal(1.00, 1)

The Eyeball Method

This eyeball method clearly has some downfalls

- Only feasible with small set of discrete values for μ
- With more parameters to estimate (ex: mean and variance) the number of candidate distributions to consider grows exponentially
- It's hard to eyeball

Better would be a disciplined measurement of "how different are the datasets?".

Maximum Likelihood Method

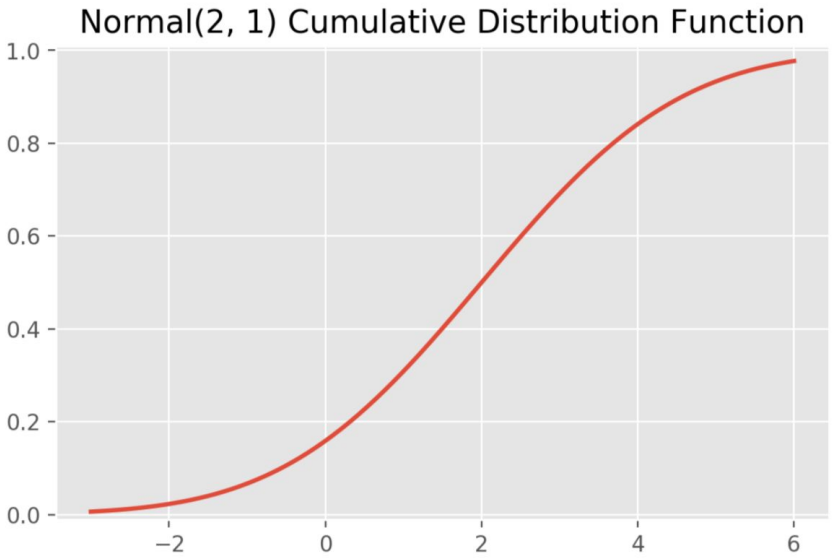
The **maximum likelihood method** is the gold standard method for fitting statistical models to data. Almost all models used in modern times use some version of the maximum likelihood method.

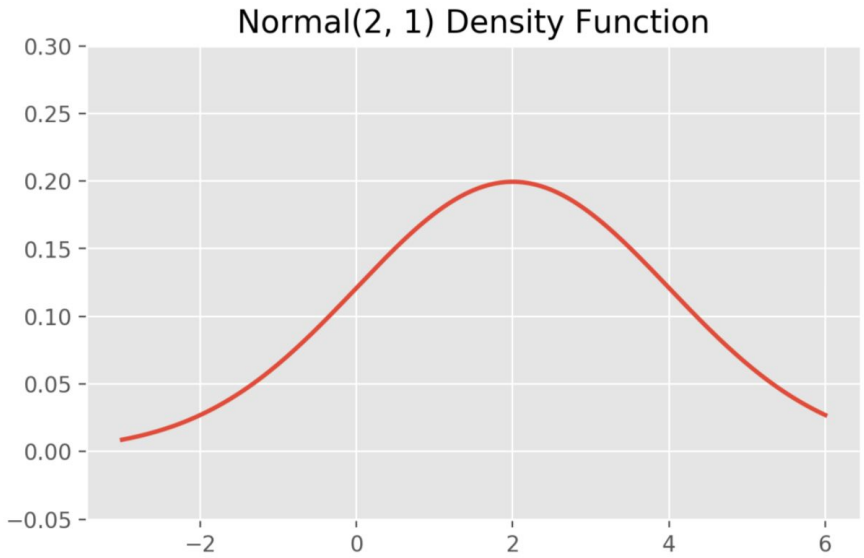
The fit model should be the random variable *most likely* to generate the data.

So how do we measure the *most likely?*

Maximum Likelihood Method

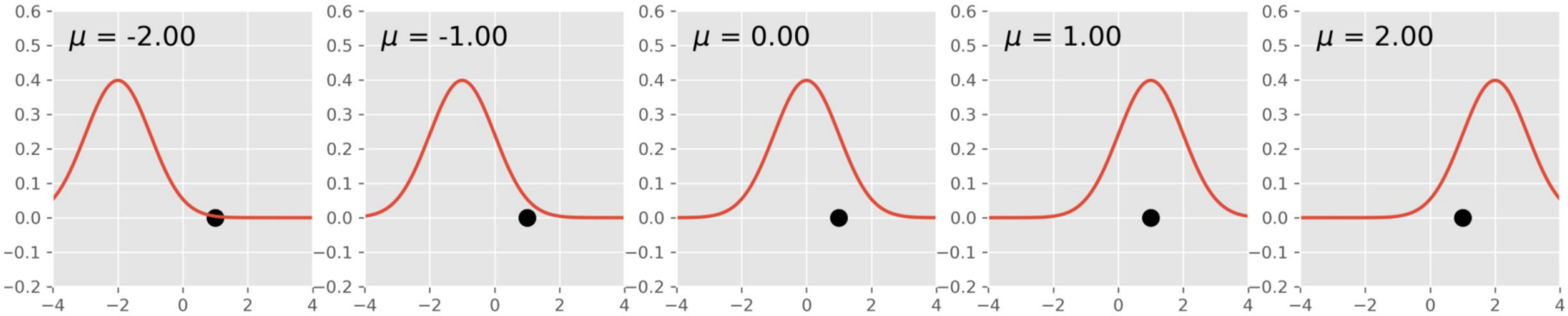
What is the data value most likely to be generated when sampling from this distribution?





Single Data Point

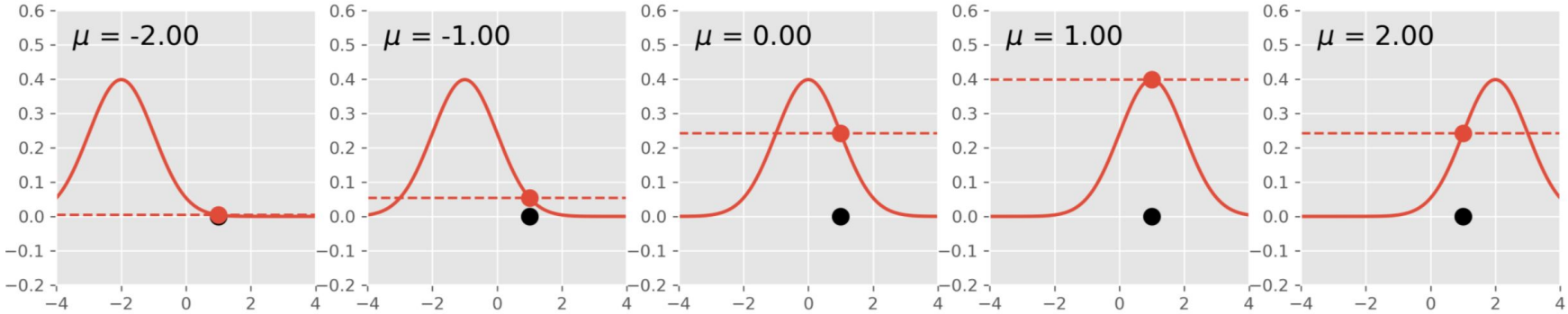
Which distribution is most likely to have generated this data point?



Single Data Point

To find the distribution that is most likely to have generated a data point x, we find the parameter θ that maximizes the density function (pdf) evaluated at x

maximize by finding θ $\{f(x;\theta)\}$

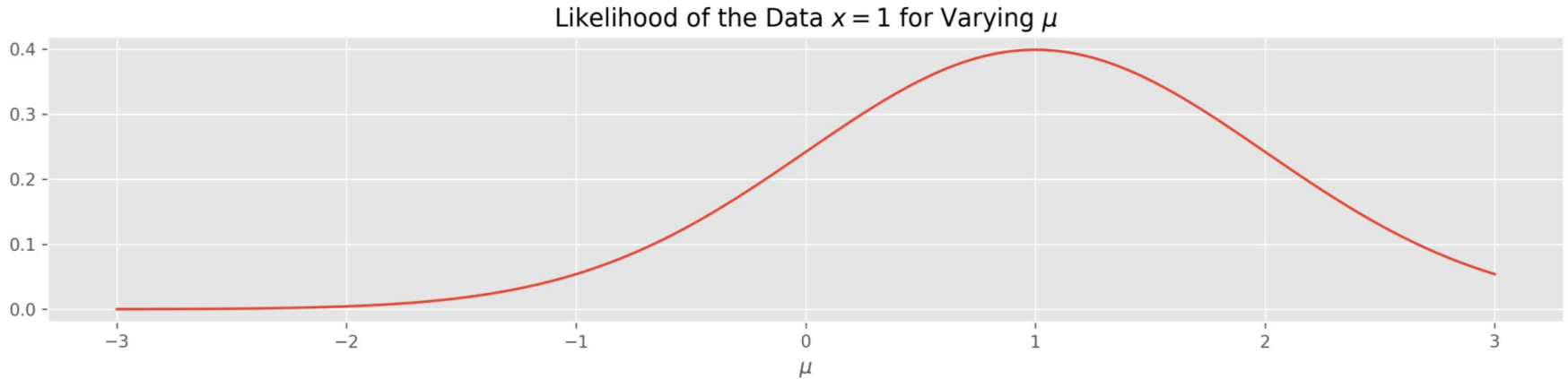


Said another way, we pick whatever value of θ maximizes the likelihood of our observed data x

Single Data Point

When our data x is fixed, and we are thinking of changing θ, we call this the likelihood function

$$L(\theta)=f(x;\theta)$$



Likelihood function

Suppose M is a statistical model, with a parameter θ. Then each of the random variables in the model has a density function, and the parameter θ appears in the density function

$$f(x_i;\theta)$$

For our the normal distribution we f(x) is (this is our pdf):

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Likelihood function

The likelihood function of the model given the data is a function of the parameter, it informally measures the likelihood of observing the data you have as you vary the parameters in the model

$$L(\theta) = \prod_i f(x_i; \theta)$$

With Data

In our case, lets image we have the following data:

$$x_1 = 0.2, \quad x_2 = -1.2, \quad x_3 = -1.5$$

Our likelihood for any given mu would be:

$$L(\mu) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(0.2-\mu)^2}{2}} \times \frac{1}{\sqrt{2\pi}} e^{-\frac{(-1.2-\mu)^2}{2}} \times \frac{1}{\sqrt{2\pi}} e^{-\frac{(-1.5-\mu)^2}{2}}$$

Easier Likelihood

In practice, the log-likelihood is more useful because it turns products into sums, and sums are easier to work with:

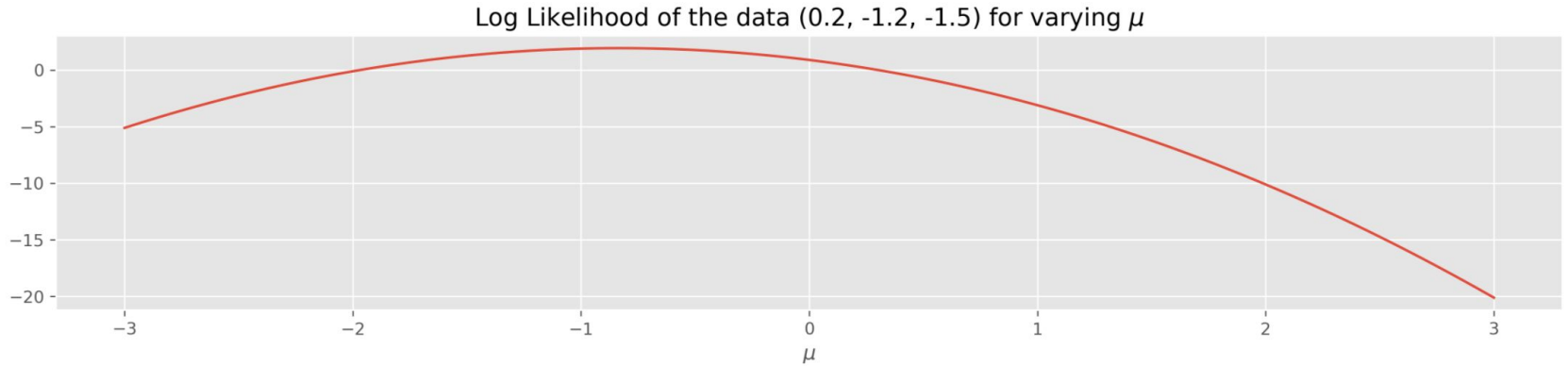
$$LL(\theta) = \sum_i \log(f(x_i, \theta))$$

In our example, using the Laws of Logarithms we get:

$$LL(\mu) = -\frac{3}{2}\log(2\pi) - \frac{(0.2-\mu)^2}{2} - \frac{(1.2-\mu)^2}{2} - \frac{(-1.5-\mu)^2}{2}$$

Fitting the Model

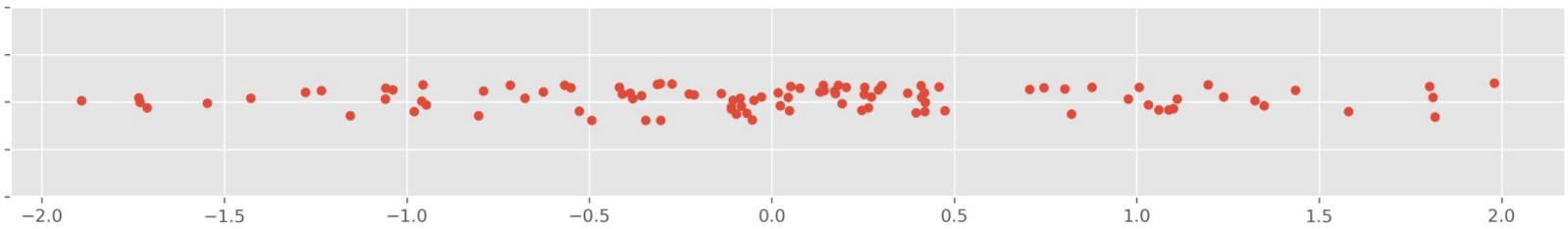
To fit a model by the likelihood method we find the parameters θ that maximize the likelihood function.



Where does the maximum log likelihood occur?

Back to our Data: Unknown Mean, STD = 1

Let’s attempt to fit a normal distribution to the data



Fitting the model to our data

Step 1: Write down the model

$$M = \{N(\mu, 1) \mid \mu \in \mathbb{R}\}$$

Step 2: Density functions of all the random variables

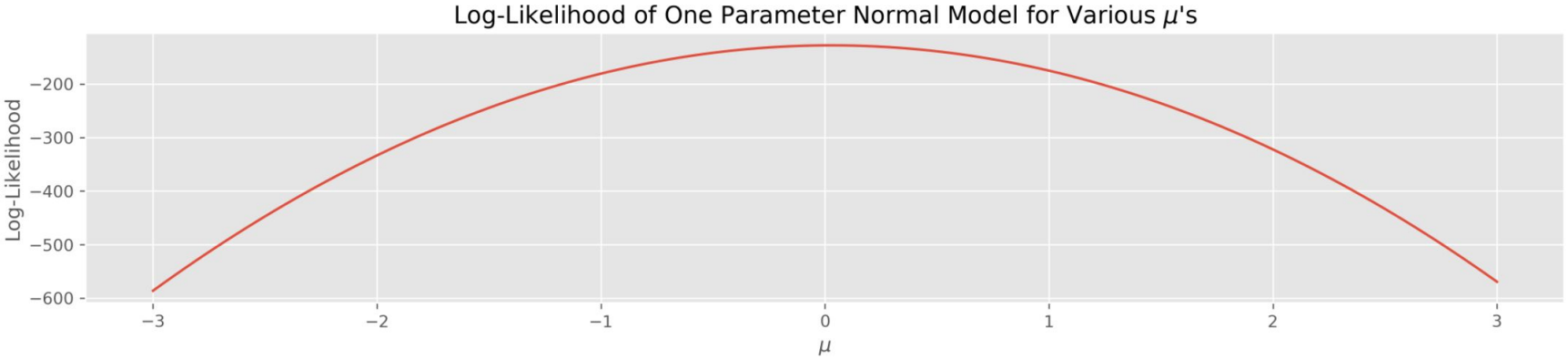
$$f(t; \mu) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(t-\mu)^2}{2}} \qquad f(\underbrace{t}_{\text{data}}; \underbrace{\mu}_{\text{parameter}})$$

Step 3: Compute the log likelihood function of the model given the data

```
def log_likelihood_normal_one_parameter(mu):
    normal = stats.norm(mu, 1.0)
    likelihoods = [normal.pdf(datum) for datum in data] # just use pdf for likelihood!
    return np.sum(np.log(likelihoods))
```

Manually

We can pass in several mu using np.linspace(-3, 3, num=250)



Step 4: Find the parameters that maximize the log-likelihood

In practice we use can either solve for the maximum, or we use gradient descent

Jupyter Notebook:

lecture-maximum-likelihood.ipynb