

Lecture 4

# Comparing Loss Functions

DSC 40A, Fall 2025

## Announcements

- Homework 1 is due on **Friday, October 10th**.
- Remember that in, general, groupwork worksheets are released on Sunday and due Monday.
- Look at the office hours schedule [here](#) and plan to start regularly attending!
- Remember to take a look at the supplementary readings linked on the course website.

# Agenda

- Recap: Empirical risk minimization.
- Choosing a loss function.
  - The role of outliers.
- Center and spread.
- Towards linear regression.

Question 🤔

Answer at [q.dsc40a.com](https://q.dsc40a.com)

Remember, you can always ask questions at [q.dsc40a.com](https://q.dsc40a.com)!

# Recap: Empirical risk minimization

## Goal

We had one goal in Lectures 2 and 3: given a dataset of values from the past, **find the best constant prediction** to make.

$$y_1 = 72 \quad y_2 = 90 \quad y_3 = 61 \quad y_4 = 85 \quad y_5 = 92$$

**Key idea:** Different definitions of "best" give us different "best predictions."

# The modeling recipe

In Lectures 2 and 3, we made two full passes through our "modeling recipe."

1. Choose a model.

$$H(x) = h$$

2. Choose a loss function.

$$L_{\text{sq}}(y_i, h) = (y_i - h)^2 \qquad L_{\text{abs}}(y_i, h) = |y_i - h|^2$$

3. Minimize average loss to find optimal model parameters.

$$h^* = \text{mean}(y_1, \dots, y_n) \qquad h^* = \text{median}(y_1, \dots, y_n)$$

# Empirical risk minimization

- The formal name for the process of minimizing average loss is **empirical risk minimization**.
- Another name for "average loss" is **empirical risk**.
- When we use the squared loss function,  $L_{\text{sq}}(y_i, h) = (y_i - h)^2$ , the corresponding empirical risk is mean squared error:

$$R_{\text{sq}}(h) = \frac{1}{n} \sum_{i=1}^n (y_i - h)^2$$

- When we use the absolute loss function,  $L_{\text{abs}}(y_i, h) = |y_i - h|$ , the corresponding empirical risk is mean absolute error:

$$R_{\text{abs}}(h) = \frac{1}{n} \sum_{i=1}^n |y_i - h|$$



## Empirical risk minimization, in general

Key idea: If  $L(y_i, h)$  is any loss function, the corresponding empirical risk is:

$$R(h) = \frac{1}{n} \sum_{i=1}^n L(y_i, h)$$

## Question 🤔

Answer at [q.dsc40a.com](https://q.dsc40a.com)

$$R_{\text{sq}}(h) = \frac{1}{n} \sum_{i=1}^n (y_i - h)^2$$

$$R_{\text{abs}}(h) = \frac{1}{n} \sum_{i=1}^n |y_i - h|$$

Is the following statement true, for any dataset  $y_1, y_2, \dots, y_n$  and prediction  $h$ ?

$$(R_{\text{abs}}(h))^2 = R_{\text{sq}}(h)$$

- A. It's true for any  $h$  and any dataset.
- B. It's true for at least one  $h$  for any dataset, but not in general.
- C. It's never true.

# Choosing a loss function

## Now what?

- We know that, for the constant model  $H(x) = h$ , the **mean** minimizes mean squared error.
- We also know that, for the constant model  $H(x) = h$ , the **median** minimizes mean absolute error.
- How does our choice of loss function impact the resulting optimal prediction?

## Comparing the mean and median

- Consider our example dataset of 5 commute times.

$$y_1 = 72 \qquad y_2 = 90 \qquad y_3 = 61 \qquad y_4 = 85 \qquad y_5 = 92$$

- As of now, the **median** is 85 and the **mean** is 80.

- What if we add 200 to the largest commute time, 92?

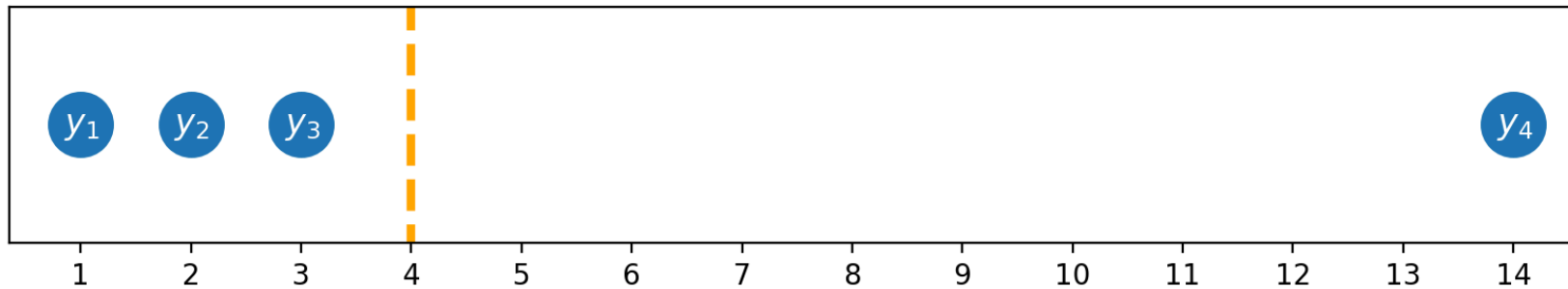
$$y_1 = 72 \qquad y_2 = 90 \qquad y_3 = 61 \qquad y_4 = 85 \qquad y_5 = 292$$

- Now, the median is 10 but the mean is 12.5!

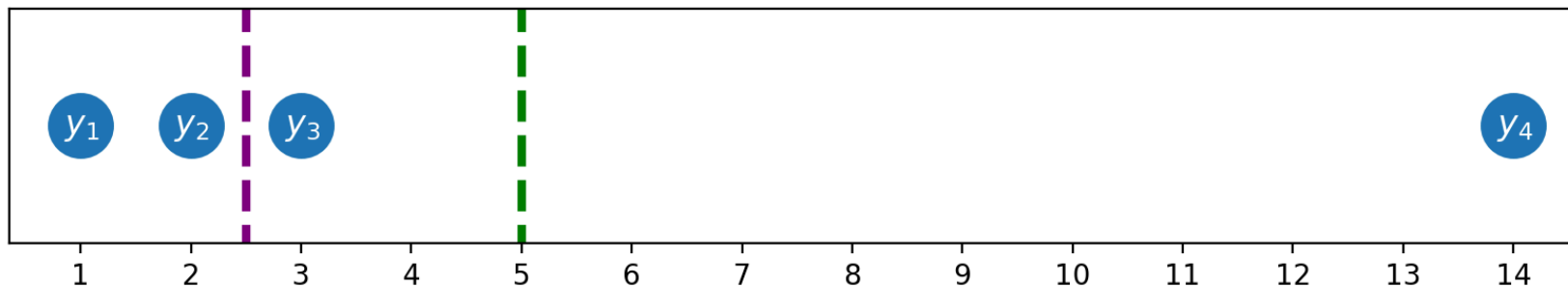
- **Key idea:** The mean is quite **sensitive** to outliers.

# Outliers

Below,  $|y_4 - h|$  is 10 times as big as  $|y_3 - h|$ , but  $(y_4 - h)^2$  is 100 times  $(y_3 - h)^2$ .

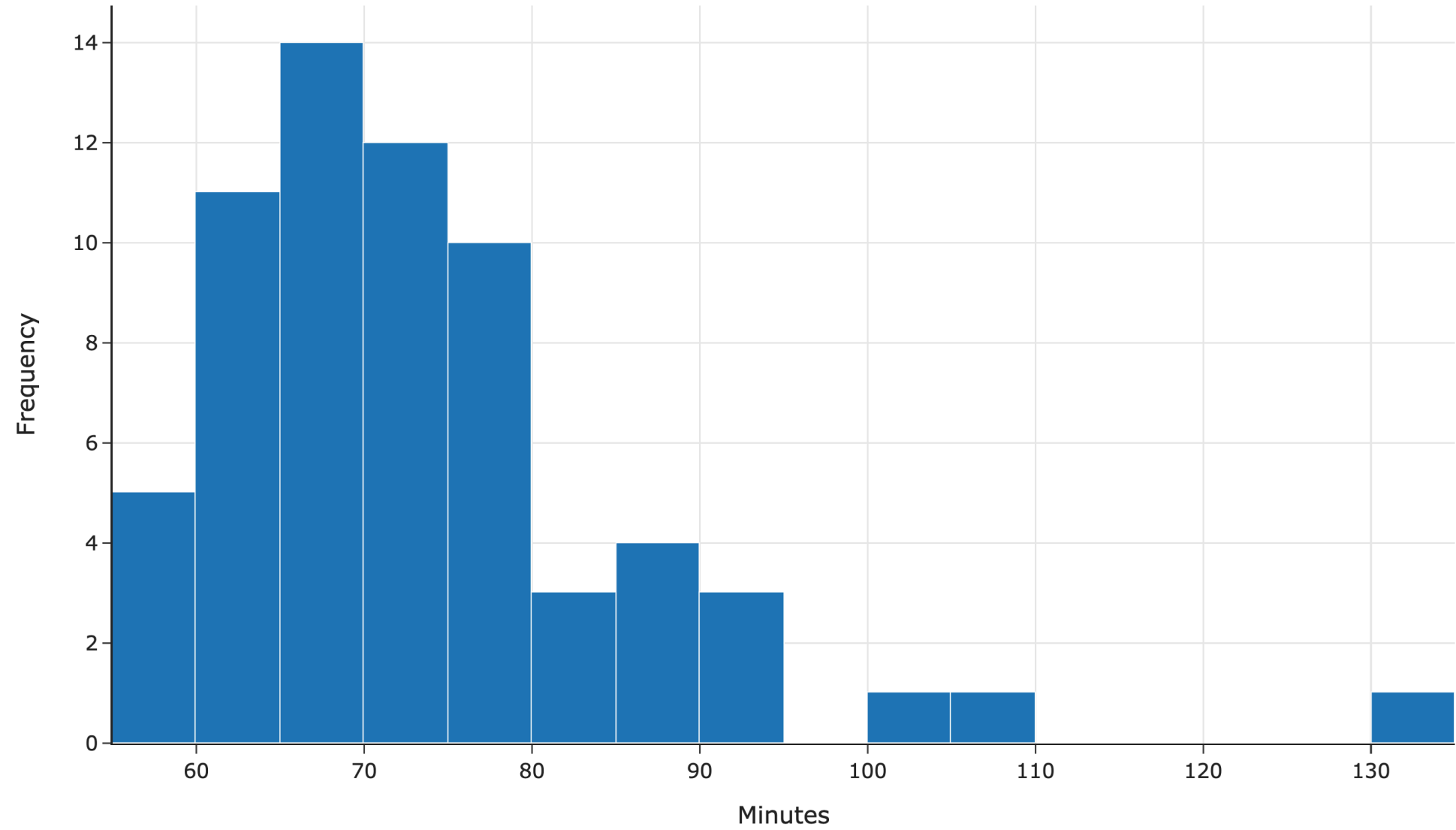


The result is that the **mean** is "pulled" in the direction of outliers, relative to the **median**.



As a result, we say the **median** is **robust** to outliers. But the **mean** was easier to solve for.

Distribution of Commuting Time

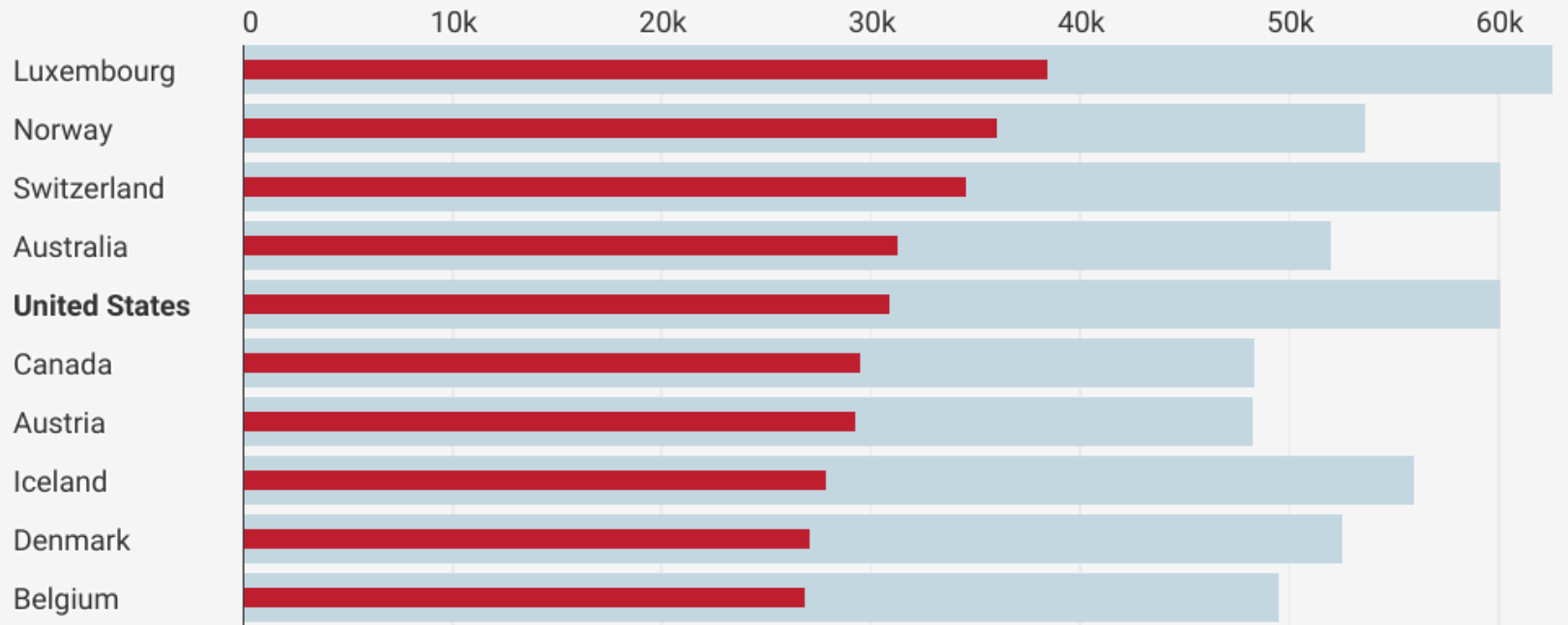


# Example: Income inequality

## Average vs median income

Median and mean income between 2012 and 2014 in selected OECD countries, in USD; weighted by the currencies' respective [purchasing power](#) (PPP).

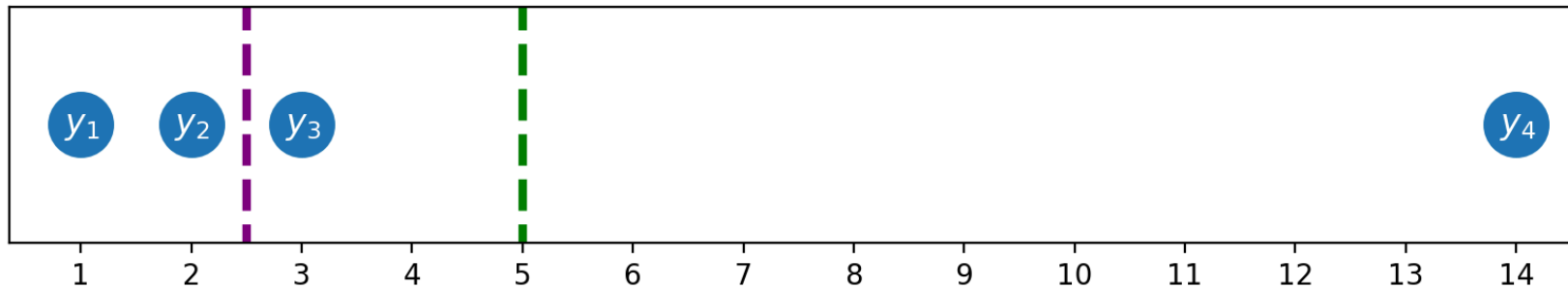
■ Average income in USD ■ Median income





## Balance points

Both the **mean** and **median** are "balance points" in the distribution.



- The **mean** is the point where  $\sum_{i=1}^n (y_i - h) = 0$ .
- The **median** is the point where  $\# (y_i < h) = \# (y_i > h)$ .

## Why stop at squared loss?

Empirical Risk, $R(h)$	Derivative of Empirical Risk, $\frac{d}{dh} R(h)$	Minimizer
$\frac{1}{n} \sum_{i=1}^n  y_i - h $	$\frac{1}{n} \left( \sum_{y_i < h} 1 - \sum_{y_i > h} 1 \right)$	median
$\frac{1}{n} \sum_{i=1}^n (y_i - h)^2$	$\frac{-2}{n} \sum_{i=1}^n (y_i - h)$	mean
$\frac{1}{n} \sum_{i=1}^n  y_i - h ^3$		???
$\frac{1}{n} \sum_{i=1}^n (y_i - h)^4$		???
$\frac{1}{n} \sum_{i=1}^n (y_i - h)^{100}$		???
...	...	...

## Generalized $L_p$ loss

For any  $p \geq 1$ , define the  $L_p$  loss as follows:

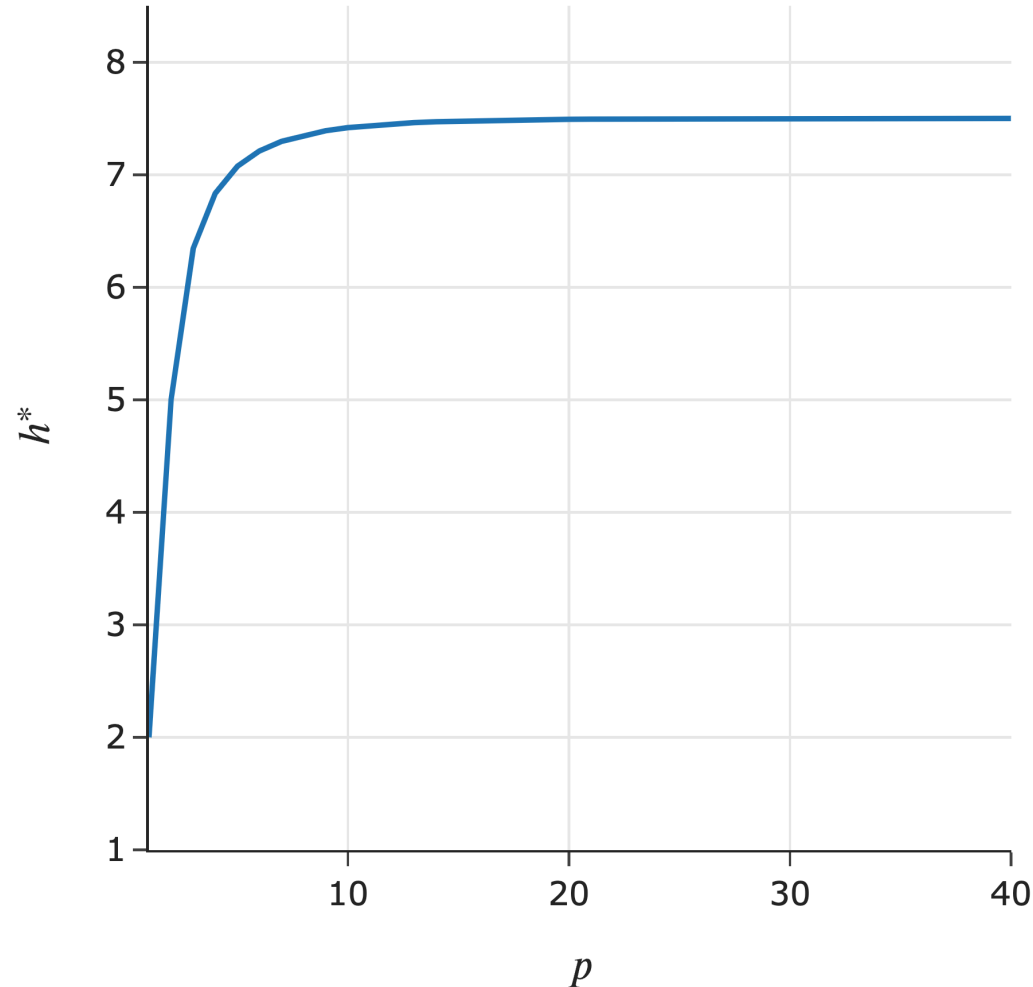
$$L_p(y_i, h) = |y_i - h|^p$$

The corresponding empirical risk is:

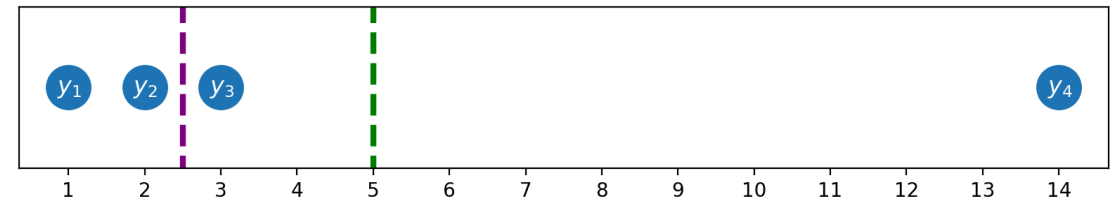
$$R_p(h) = \frac{1}{n} \sum_{i=1}^n |y_i - h|^p$$

- When  $p = 1$ ,  $h^* = \text{Median}(y_1, y_2, \dots, y_n)$ .
- When  $p = 2$ ,  $h^* = \text{Mean}(y_1, y_2, \dots, y_n)$ .
- What about when  $p = 3$ ?
- What about when  $p \rightarrow \infty$ ?

What value does  $h^*$  approach, as  $p \rightarrow \infty$ ?



Consider the dataset 1, 2, 3, 14:



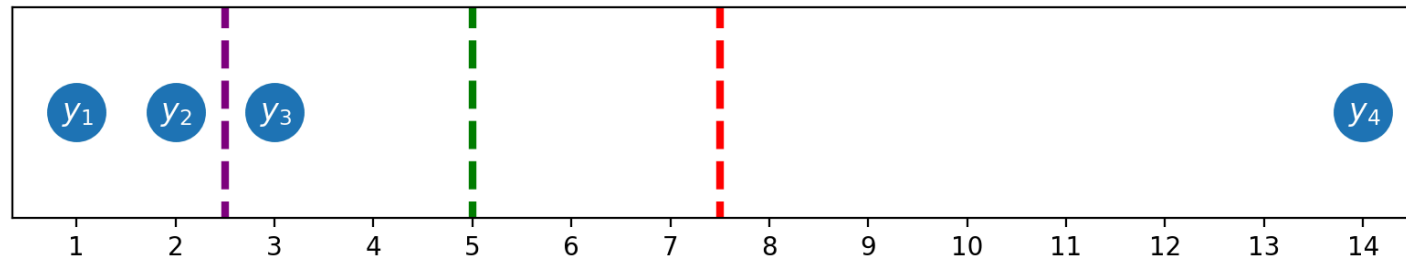
On the left:

- The  $x$ -axis is  $p$ .
- The  $y$ -axis is  $h^*$ , the optimal constant prediction for  $L_p$  loss:

$$h^* = \operatorname{argmin}_h \frac{1}{n} \sum_{i=1}^n |y_i - h|^p$$

## The *midrange* minimizes average $L_\infty$ loss!

On the previous slide, we saw that as  $p \rightarrow \infty$ , the minimizer of mean  $L_p$  loss approached the midpoint of the minimum and maximum values in the dataset, or the **midrange**.



- As  $p \rightarrow \infty$ ,  $R_p(h) = \frac{1}{n} \sum_{i=1}^n |y_i - h|^p$  minimizes the "worst case" distance from any data point". (Read more [here](#)).
- If your measure of "good" is "not far from any one data point", then the midrange is the best prediction.

## Another example: 0-1 loss

Consider, for example, the **0-1 loss**:

$$L_{0,1}(y_i, h) = \begin{cases} 0 & y_i = h \\ 1 & y_i \neq h \end{cases}$$

The corresponding empirical risk is:

$$R_{0,1}(h) = \frac{1}{n} \sum_{i=1}^n L_{0,1}(y_i, h)$$

## Question 🤔

Answer at [q.dsc40a.com](https://q.dsc40a.com)

$$R_{0,1}(h) = \frac{1}{n} \sum_{i=1}^n \begin{cases} 0 & y_i = h \\ 1 & y_i \neq h \end{cases}$$

Suppose  $y_1, y_2, \dots, y_n$  are all unique. What is  $R_{0,1}(y_1)$ ?

- A. 0.
- B.  $\frac{1}{n}$ .
- C.  $\frac{n-1}{n}$ .
- D. 1.













## Minimizing empirical risk for 0-1 loss

$$R_{0,1}(h) = \frac{1}{n} \sum_{i=1}^n \begin{cases} 0 & y_i = h \\ 1 & y_i \neq h \end{cases}$$



## Summary: Choosing a loss function

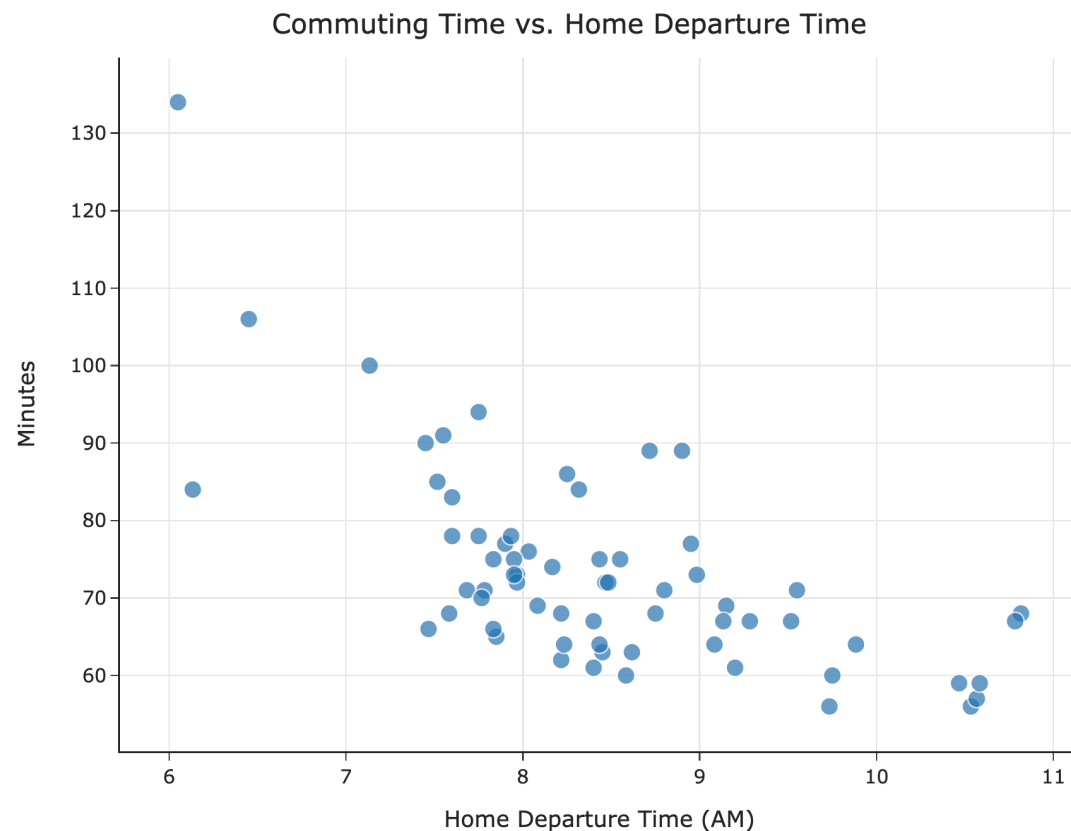
Key idea: Different loss functions lead to different best predictions,  $h^*$ !

Loss	Minimizer	Always Unique?	Robust to Outliers?	Differentiable?
$L_{sq}$	mean	yes 	no 	yes 
$L_{abs}$	median	no 	yes 	no 
$L_{\infty}$	midrange	yes 	no 	no 
$L_{0,1}$	mode	no 	yes 	no 

The optimal predictions,  $h^*$ , are all **summary statistics** that measure the **center** of the dataset in different ways.

What's next?

# Towards simple linear regression



- In Lecture 1, we introduced the idea of a hypothesis function,  $H(x)$ .
- We've focused on finding the best **constant model**,  $H(x) = h$ .
- Now that we understand the modeling recipe, we can apply it to find the best **simple linear regression model**,  $H(x) = w_0 + w_1x$ .
- This will allow us to make predictions that aren't all the same for every data point.

# The modeling recipe

1. Choose a model.
2. Choose a loss function.
3. Minimize average loss to find optimal model parameters.

