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.
- Other loss functions

Question 🤔

Answer at q.dsc40a.com

Remember, you can always ask questions at 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.

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

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

mean → median →

both are the "best" under different loss functions

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:

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

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

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

Empirical risk minimization, in general

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

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

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?**

$$\frac{72+90+61+85+92}{5} = \frac{400}{5}$$

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?

+200

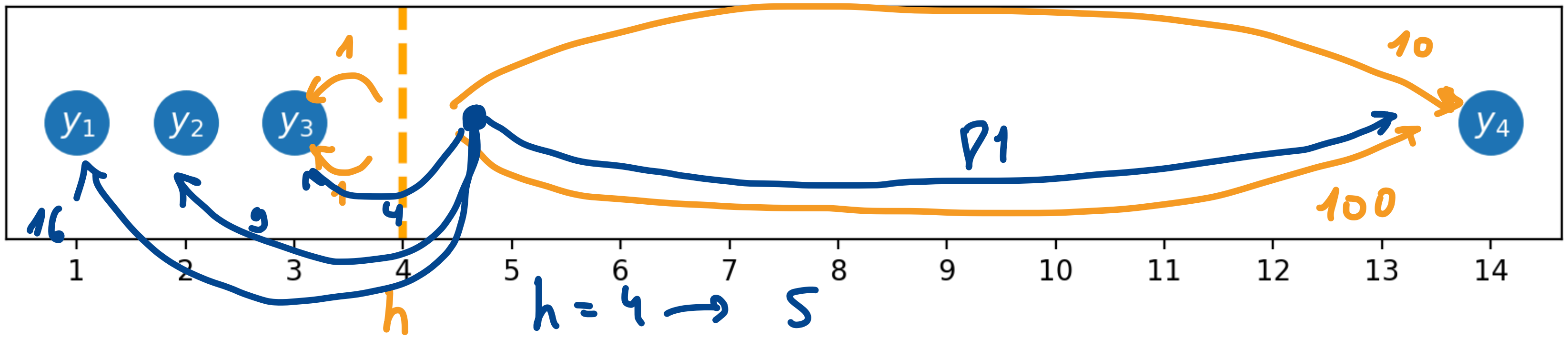
$$y_1 = 72 \qquad y_2 = 90 \qquad y_3 = 61 \qquad y_4 = 85 \qquad y_5 = 292$$

- Now, the median is still 85 but the mean is 120 !
- **Key idea**: The mean is quite **sensitive** to outliers.

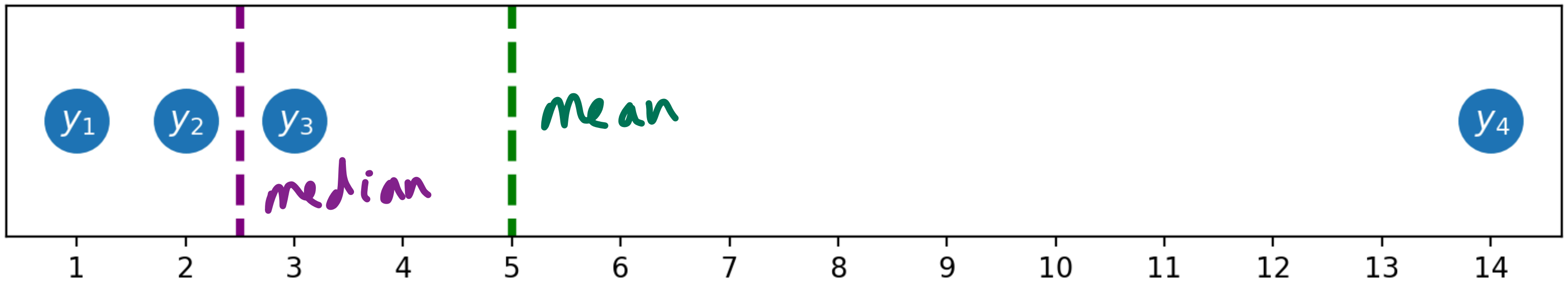
$$80 + \frac{200}{5} = 120$$

Outliers

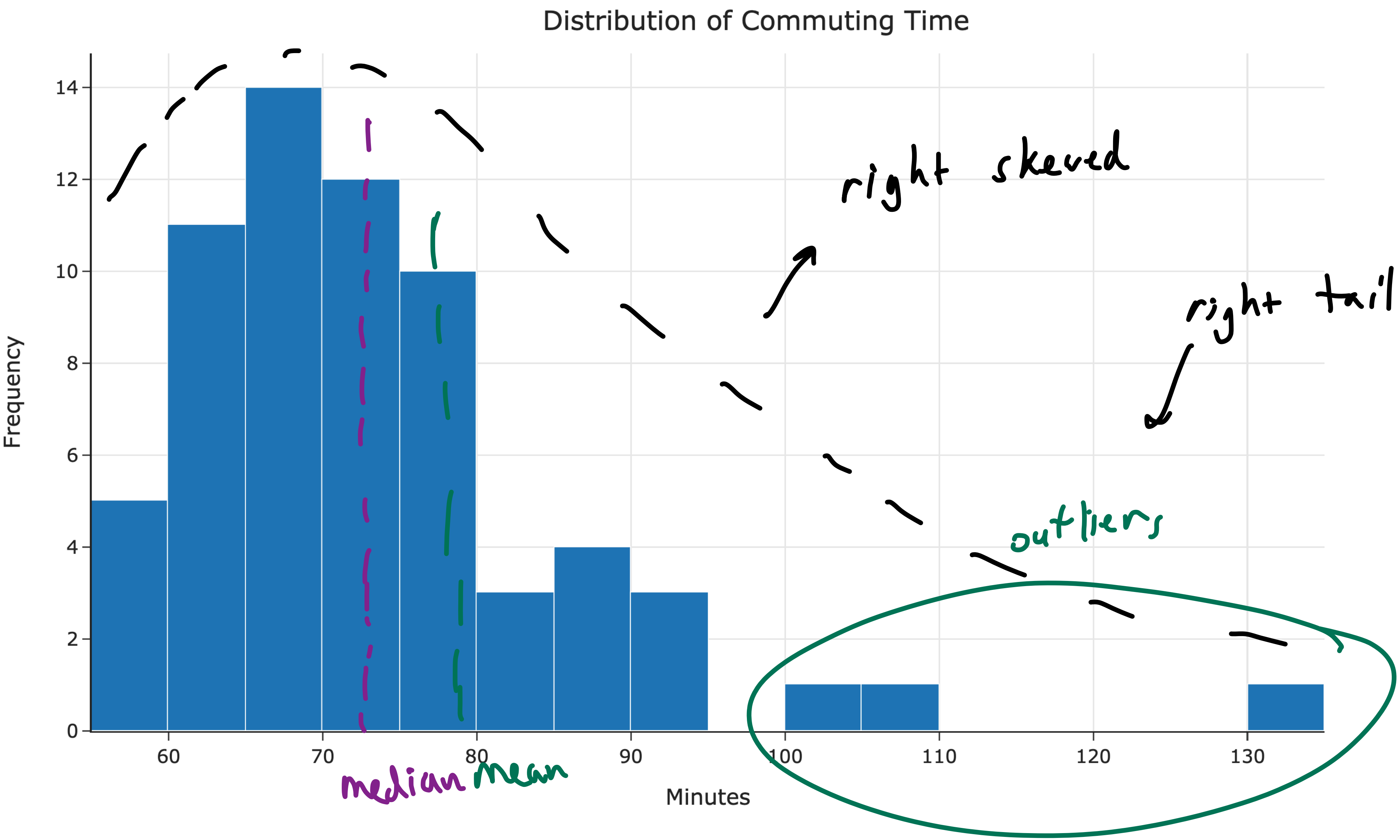
Below, $|y_4 - h|$ is 10 times as big as $|y_3 - h|$, but $(y_4 - h)^2$ is 100 times $(y_3 - h)^2$.



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

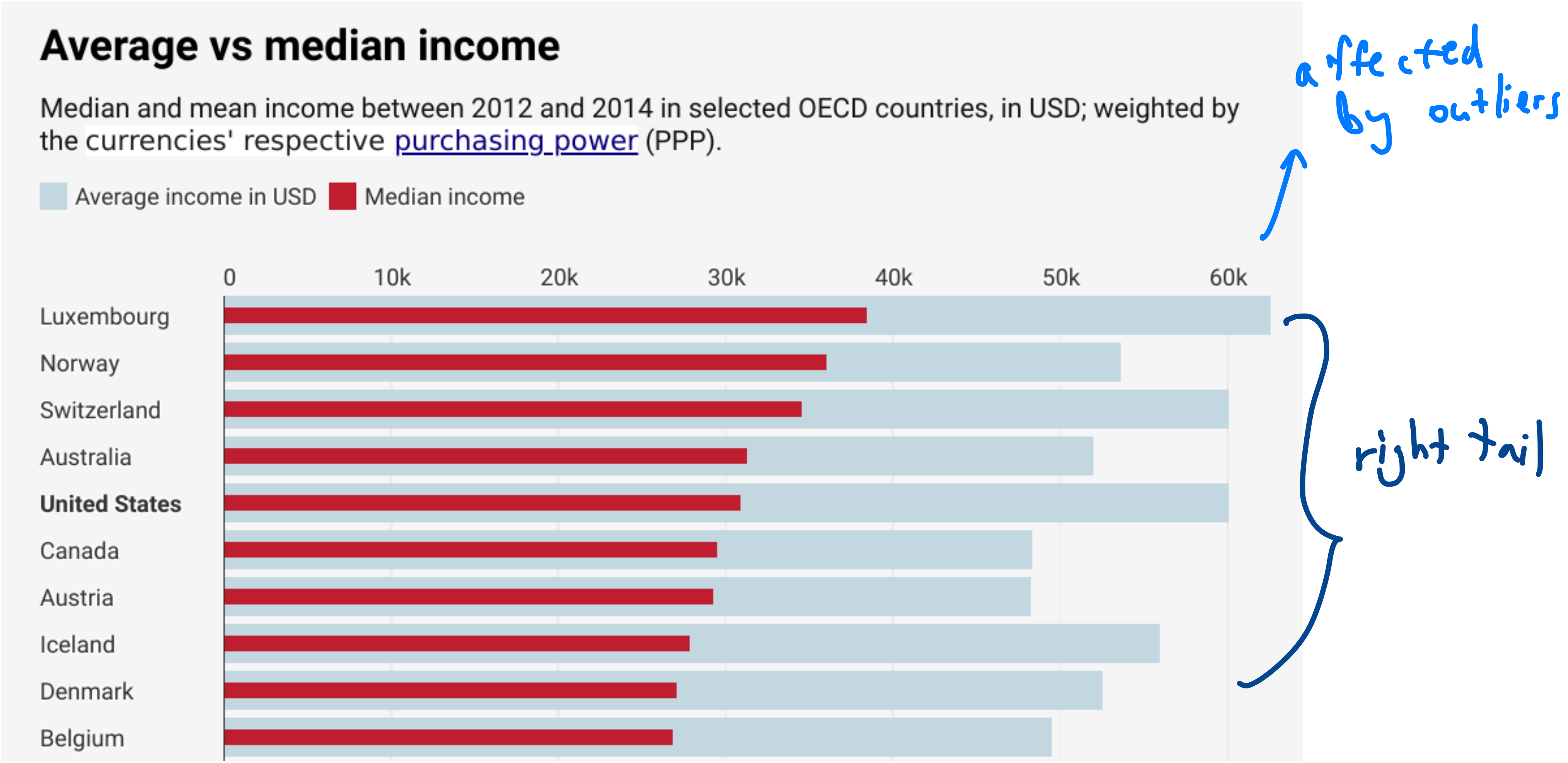


As a result, we say the **median** is **robust** to outliers. But the **mean** was easier to solve for.



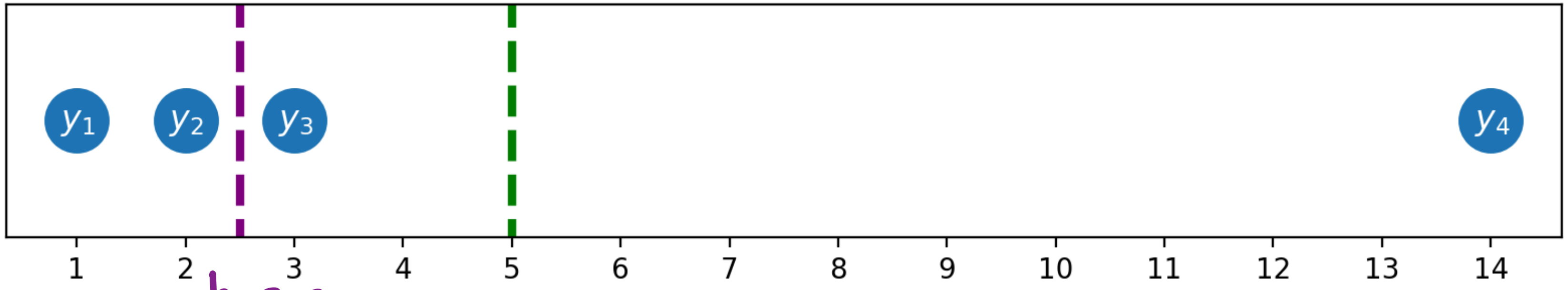
Distribution of Commuting Time
Frequency
Minutes
0
2
4
6
8
10
12
14
60
70
80
90
100
110
120
130
right skewed
right tail
outliers
median
mean

Example: Income inequality



Balance points

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



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

$\frac{d}{dh}\sum_{i=1}^{n}(y_i - h)^2$

$\sum_{y_i < h^*} |y_i - h^*| = \sum_{y_i > h^*} |y_i - h^*|$

- The **median** is the point where $\#\,(y_i < h) = \#\,(y_i > h)$.

we have 2 data points left of $h^* = 2.5$ (y_1, y_2)
2 " right of $h^* = 2.5$ (y_3, y_4)

Why stop at squared loss?

Empirical Risk, $R(h)$	Derivative of Empirical Risk, $\frac{d}{dh}R(h)$	Minimizer
$\frac{1}{n}\sum_{i=1}^{n}\lvert y_i - h\rvert$	$\frac{1}{n}\left(\sum_{y_i<h} 1 - \sum_{y_i>h} 1\right) = 0$	median
$\frac{1}{n}\sum_{i=1}^{n}(y_i - h)^2$	$\frac{-2}{n}\sum_{i=1}^{n}(y_i - h) = 0$	mean
$\frac{1}{n}\sum_{i=1}^{n}\lvert y_i - h\rvert^3$		???
$\frac{1}{n}\sum_{i=1}^{n}(y_i - h)^4$	$-\frac{4}{n}\sum_{i=1}^{n}(y_i - h)^3 = 0$ →	???
$\frac{1}{n}\sum_{i=1}^{n}(y_i - h)^{100}$		???
...	...	...

* $(y_i - h)^3$ could be negative

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

p-norm of a vector

$$||\vec{x}||_2 = \sqrt{x_1^2 + x_2^2 + \ldots + x_n^2}$$

$$||\vec{x}||_3 = \sqrt[3]{x_1^3 + x_2^3 + \ldots + x_n^3}$$

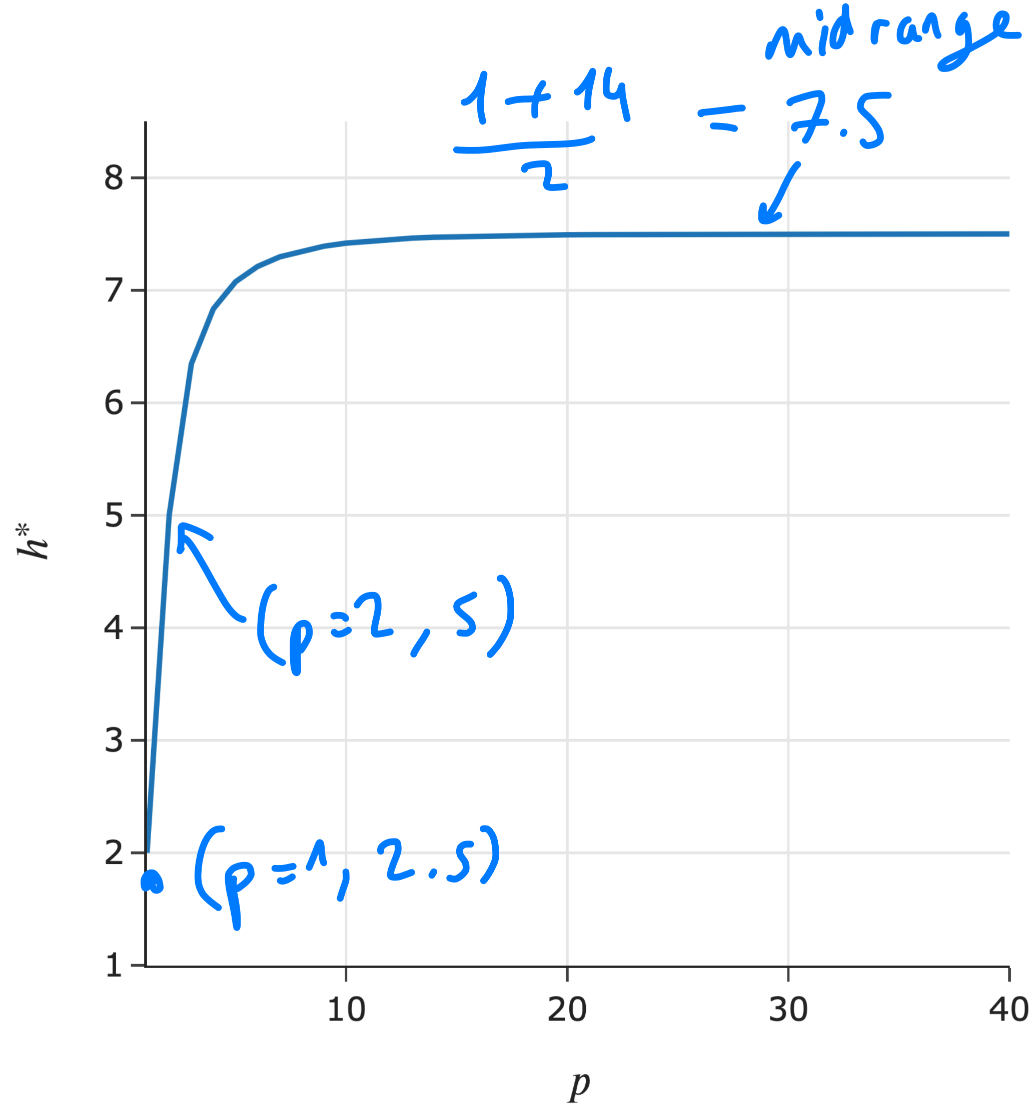
$$\vdots$$

$$||\vec{x}||_{100} = \sqrt[100]{x_1^{100} + x_2^{100} + \ldots + x_n^{100}}$$

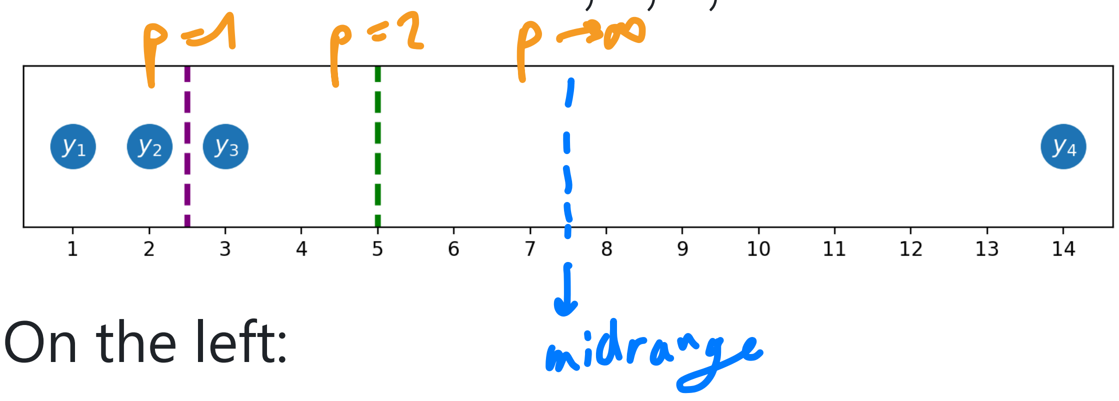
$$||\vec{x}||_\infty = \max(x_1, \ldots, x_n)$$

$$||\vec{x}||_p = \sqrt[p]{x_1^p + \ldots + x_n^p}$$

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



Consider the dataset $1, 2, 3, 14$:



On the left:

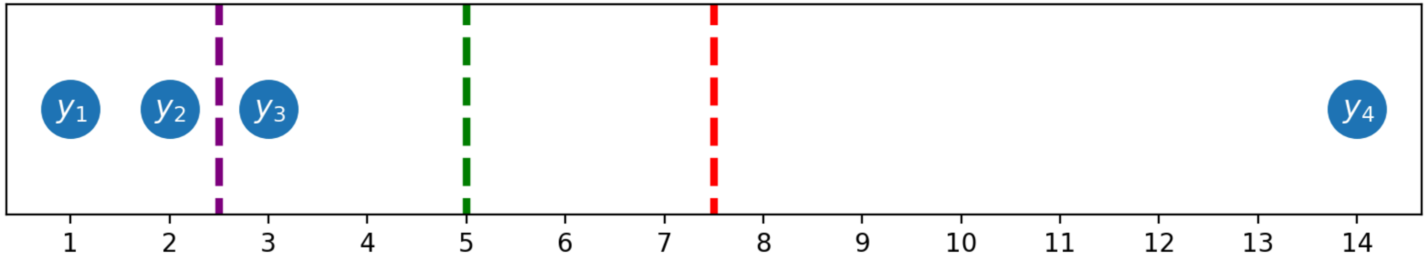
- The x-axis is p.
- The y-axis is h^*, the optimal constant prediction for L_p loss:

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

"infinity" loss

The *midrange* minimizes average L_∞ loss!

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



- As $p \to \infty$, $R_p(h) = \frac{1}{n}\sum_{i=1}^{n} |y_i - h|^p$ 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.

mean = 5, worst case distance |14 − 5| = 9
median = 2.5 worst case distance |14 − 2.5| = 11.5
midrange = 7.5 worst case distance |14 − 7.5| = |7.5 − 1| = 6.5